Matej Bel University, Banská Bystrica, Slovakia
Has been issued since 2014
ISSN 1339-6773
E-ISSN 1339-875X

Forecasting the Price Index Return and Movement Direction using Data Mining Techniques

[1] Günter Şenyurt
[2] Abdülhamit Subaşı

[1] International Burch University, Bosnia and Herzegovina
Francuske revolucije bb, Ilidza 71210
M. Sc. (Information Technologies), Assistant Lecturer
E-mail: gunter.senyurt@ibu.edu.ba
[2] International Burch University, Bosnia and Herzegovina
Francuske revolucije bb, Ilidza 71210
Dr. (Electrical Eng.), Professor
E-mail: abdulhamit.subasi@ibu.edu.ba

Abstract

Even though many new data mining techniques have been introduced in prediction estimation, there is still no single best solution to all financial problems. In this study, the performances of data mining techniques based on the daily Istanbul Stock Exchange (ISE) Index are examined and compared. The linear regression model, simple logistic (classification), artificial neural networks (ANN) and support vector machines (SVM) models are utilized in two ways, one for classification of market movements and the other for predicting price index returns through regression. Ten technical market indicators, 7 macroeconomic variables, a couple of other international market indices and a sliding window of ten inputs make up the 30 attributes used in this study. Different combinations of attribute sets are experimented with different ANN and SVM model parameter values to find the highest forecasting accuracy.

Keywords: ANN, Data Mining Techniques, Forecasting, Market movement direction, Price index return, SVM.

Introduction

It is of utmost importance for investors to estimate the trend of the markets as precisely as possible in order to reach the best trading decisions for their investments, so in this context it is in the investor's best interest to use the most accurate time series forecasting model to maximize the profit or to minimize the risk. All in all, it is a quite challenging job to make accurate predictions of stock market index movements and model the time series data, especially in highly volatile markets such as the Turkish stock market. That is due to the fact that stock markets are in general chaotic and complex mechanisms with dynamic, nonlinear and nonparametric variables [1]. Moreover, markets are influenced by numerous macroeconomic factors, institutional investor choices, human psychology, political events, company policies, other stock market movements and economic affairs [2]. In this study it is intended to introduce several time series prediction models such as linear regression, simple logistic, artificial neural network (ANN), support vector machines (SVM) and compare their performance based on the daily Istanbul Stock Exchange (ISE) data. There is lots of empirical work available in literature on well-established and developed markets such as Dow Jones (USA) or DAX (Germany), whereas little research material is available on new emerging markets such as ISE [3]. By means of this study, it is aimed at contributing to the demonstration and verification of the XU-100 index price level predictability through a number of time series forecasting regression models whose names were mentioned earlier above. The related predicting performances of these models are compared based on statistical criteria such as relative absolute error (RAE), root relative squared error (RRSE) and the squared value of the correlation coefficient

(R^2) for regression analysis. In case of classification, the percentage of accuracy is calculated and tabulated.

Literature Review

The direction of movements of a variety of financial instruments has attracted a growing number of researchers, lately, as a core subject [3]. Many academic people and professionals have put tremendous effort into forecasting stock market index future movements and figuring out a sound trading strategy that is able to turn the forecast results into profit [4]. In this section earlier studies on linear regression, ANN and SVM in financial forecasting are presented.

1. Linear Regression

It has been suggested by substantial evidence in the financial econometric literature that to some extent, excess stock market returns can be forecasted. However, several studies point out that only the direction of stock returns are predictable due to fact that the noise hidden in the observed data makes it hard to forecast the index return precisely [5]. [6] experimented with several multivariate classification methods in forecasting the direction of the index return showing that basic prediction tools such as adaptive exponential smoothing and vector auto regression with Kalman filter updating were outperformed by other classification models such as logit, discriminant analysis and probit methods [3]. The auto-logistic model was used by [7] and [8] to forecast the direction of returns while [9] suggested a new dynamic probit model to be employed in the directional predictability of stock market returns [5]. In "Forecasting the direction of the US stock market with dynamic probit models", the results show the probit models' statistical significance of "in-sample predictive power for excess stock market return signs" [5]. The Ordinary Least Square (OLS) regression technique was compared by [10] with their neural network model in predicting the ISE-30 and ISE-ALL indices showing that the neural network model has potential to predict better than the linear regression model.

2. Artificial Neural Networks (ANN)

There are various ANN methods that can be used in predicting stock price returns and movement directions and a great deal of research has been conducted on using ANN to forecast financial time series data outputs suggesting ANN as a powerful tool in predicting stock market return [11] and [12]. [4] used the probabilistic neural network (PNN) which showed strong predictive power over other models such as the GMM-Kalman filter and random walk. [13], who trained back propagation neural networks, based the input attributes on some technical market indicators like momentum, moving average, moving average convergence divergence (MACD), RSI and stochastic %K and forecasted the ISE 100 index direction with % 60.81 accuracy while [10] also used ISE-30 and ISE-ALL indices to see the performances of several neural network models. [14] effectively proved that multivariate neural networks could outperform the linear models for stock price movement predictions of Shanghai Stock Exchange listed companies.

3. Support Vector Machines (SVM)

The support vector machines technique has proved to be a promising new technique, lately, in stock price index movement directions and stock price return forecasting. SVM was used by [15] to experiment the daily stock price change in KOSPI (Korean Stock Price Index). Using 12 technical indicators such as momentum, stochastic %K, stochastic %D, RSI, A/D oscillator and ROC, the feasibility of SVM in market forecasting was tested along with back-propagation (BPN) networks and case-base reasoning (CBR). The result showed the potential of SVM in correctly predicting the output even better than BPN and CBR. In [16] the traditional discriminant, logit models and ANN was compared with SVM and random forest to examine results with S&P CNX NIFTY market index of the National Stock Exchange. They used the same attributes as done by [15] and SVM proved more powerful than the other techniques. [17] based their experiment on the NIKKEI 225 index using SVM, linear discriminant analysis, quadratic discriminant analysis and Elman BPN. They found that the weekly movement direction of NIKKEI 225 could be more accurately predicted by the SVM classification method in comparison with the other techniques. In another study, [18] compared the forecasting performances of ARIMA, ANN, SVM and random forest regression techniques to find that SVM outperformed the other models used in the experiment. They also

developed a model with a two-stage architecture where they integrated a self-organizing map and a support vector regression to examine several major stock market indices. The results proved that the two-stage model could be used as an alternative in market price forecasting. [3] used a three-layered feed-forward ANN structure and SVM to predict the direction of the Istanbul Stock Exchange through a dataset based on the XU-100 index from 1997 to 2007. Their BPN model predicted the movement direction with an average of 75.74% accuracy, while the SVM model result was only 71.52%, yet outperforming [13]'s and [10]'s results.

Materials and Methods

1. Research Data

In this section, the research data and the input attributes are described. The daily closing prices of the ISE National 100 Index (XU-100) covering the period from January 2, 1997 to December 31, 2007 was implemented. The total number of cases or 2733 trading days have 1440 days with increasing direction (advances), while 1293 days show decreasing direction (declines). The same dataset that was generated by the technical analysis module of Matriks gold 2.4.0, a product of Matriks Information Delivery Services Inc. and employed by [3] in their paper was integrated as part of the main dataset of this study for performance comparisons of the models used. While they only examine the direction of movement prediction performances, this study includes the return price regression results for each model, as well. All experiments were conducted on WEKA software using its Simple logistic, Linear regression, SVM and MLP built-in tools to make comparisons of prediction performances based on the chosen dataset.

The full dataset is comprised of 30 input variables in total. The first 10 in-put attributes are technical market indicators as used by [3], which are 10-day moving average, 10-day weighted moving average, momentum, stochastic %K, stochastic %D, RSI (Relative Strength Index), MACD (moving average convergence divergence), Larry William's %R, A/D (Accumulation/Distribution) Oscillator and CCI (Commodity Channel Index) which are explained shortly in the next part. Another 10 inputs are mainly chosen from macroeconomic variables, consisting of USD (sell)-Turkish Lira exchange rate, gold price (close), monthly interest rate, CPI (consumer price index), WPI (wholesale price index), PPI (producer price index), Industrial Production Index, DJI (Dow Jones) closing price, DAX (Germany) closing price and BOVESPA (Brazil) closing price. These variables are slightly differently chosen than [19]'s input variables. The final 10 inputs are a sliding window of the last 10 elements of XU-100 closing price index. In [20], an input window size of seven was used but it is preferred to use the last 10 elements in this study. The simple logistic function of WEKA (Waikato Environment for Knowledge Analysis) was utilized instead of the linear regression function in regression evaluations. For both classification and regression analysis, 10-fold cross-validation was used as the test option in WEKA (1999-2010).

2. Linear Regression Model

Linear regression is extensively used in financial forecasting which can be formulated as follows,

$$y_t = \sum \beta_k \cdot x_{k,t} + \epsilon_t \quad (1)$$

$$\epsilon_t \sim N(0, \sigma^2) \quad (2)$$

so as the variable ϵ_t is defined as a "random disturbance term" that is "normally distributed with mean zero and constant variance σ^2, and $\{\beta_k\}$ represents the parameters to be estimated." The estimated parameter set "is denoted by $\{\hat{\beta}_k\}$", while the forecast set of y which is produced "by the model with the coefficient set $\{\hat{\beta}_k\}$, is denoted by $\{\hat{y}_t\}$." The model aims "to select $\{\hat{\beta}_k\}$ such that "the sum of squared differences between the actual observations y and the observations predicted by the linear model $\hat{y}$ is minimized [21].

The time series input and output variables, [y x], use subscript t indicating the particular observation date, with observations starting at t=1. Various methods are available for estimating the parameter set $\{\beta_k\}$, with many alternative assumptions made on the distribution of the disturbance term, ϵ_t, and the constancy of its variance, σ^2. The independence of the distribution of

the input variables x_k with respect to the disturbance term, ϵ_t, can also be estimated by certain assumptions. In the linear regression estimation process it is aimed to find a set of parameters for the model given by $\{\hat{\beta}_k\}$, in order to minimize Ψ, that is described as the sum of squared differences, errors or residuals, between the target (observed or output) value y and the model predicted variable $\hat{y}$ [21]. The problem of estimation can be expressed in the following way:

$$Min_{\hat{\beta}}\Psi = \sum_{t=1}^{T}\hat{\epsilon}_t^2 = \sum_{t=1}^{T}(y_t - \hat{y}_t)^2 \quad (3)$$

given that

$$y_t = \sum \beta_k \cdot x_{k,t} + \epsilon_t \quad (4)$$

$$\hat{y}_t = \sum \hat{\beta}_k \cdot x_{k,t} \quad (5)$$

$$\epsilon_t \sim N(0, \sigma^2) \quad (6)$$

As a tool of forecasting, the autoregressive linear model is utilized as follows:

$$y_t = \sum_{i=1}^{k^*}\beta_i \cdot y_{t-1} + \sum_{j=1}^{k}\gamma_j \cdot x_{j,t} + \epsilon_t \quad (7)$$

so as there are k independent $x$variables with coefficient $\gamma_j$for each x_j and k^* lags for the dependent variable y and $k + k^*$ parameters, {β} and {γ}, are to be estimated [21].

3. Artificial Neural Network (ANN) Model

Artificial neural networks are capable estimation models for financial modeling and prediction [3]. In this study, a three layered feed-forward ANN structure (a multilayer perceptron) is used to forecast stock market index movements. Multilayer perceptrons (MLP) have one or more layers between input and output layers, called hidden layers that can approximate any nonlinear relation to any accuracy given sufficiently large number of neurons. The nonlinearity used in the nodes provides MLP with a universal approximation power. “It has been scientifically proved that a three-layered MLP using sigmoidal activation function can approximate well any continuous multivariate function to any accuracy” [22]. MLP shows high efficiency in function approximation for high-dimensional spaces. It has clear advantage over linear regression methods in that the input dimensionality does not affect the error convergence rate, while conventional linear regression methods suffer from the size of dimensionality. The most popular learning rule in supervised learning is the back propagation learning algorithm which is used to train the neural network. In order to minimize a cost function that is equivalent to MSE (mean squared error) between the desired and actual network outputs, a gradient search method is utilized. An input pattern is introduced to the system and the resulting computed output is compared with the actual given output (target output). The error of each calculated output is propagated backward that establishes a closed-loop control system which adjusts weights by a gradient-descend based algorithm [22]. Neural networks were initially derived as models representing the human brain. Each unit is represented as a neuron while the connections (links) represent synapses and in early neural network models, when the total signal passed to a unit exceeds a certain threshold the neurons fired. In earlier models, this concept was adopted using a step function as a threshold function for nonlinear statistical modeling, though later it was replaced by the sigmoid function for smoother optimization. The unknown parameters of the neural network are called weights which are sought to make the model fit the training data well.

For regression, the sum of squared errors could be used as a measure of fit (error function). For classification, the squared error as well as cross-entropy (deviance) can be used for a fit or error function [23]. Neural networks happen to have too many weights that overfit the data at the global minimum of R: In early models the designers introduced an early stopping rule where the model is trained only for a while before actually reaching the global minimum in order to avoid the overfitting problem. Weights are generally started out at a highly regularized (linear) solution having the effect of compressing the final model toward a linear one. In this case, a validation dataset is used to determine when to stop due to the fact that the validation error is expected to start growing. The effective scaling of the weights in the bottom layer is affected by the scaling of the inputs which is having a direct influence on the final result. The number of hidden layers is chosen by experimentation and background knowledge, but the range usually differs between 5 to 100 that increases with the number of inputs and number of training cases. Cross validation is a useful tool to either estimate the optimal number of hidden layers or the regularization parameters. Each layer can extract features of the input attributes for both classification and regression.

The use of multiple hidden layers is also possible to construct hierarchical features for several levels of resolution [23].

4. Support Vector Machine (SVM) Model

SVM is implemented with the structural risk minimization principle that is found in statistical learning theory [22]. "Structural risk minimization (SRM) is an inductive principle of use in machine learning. Commonly in machine learning, a generalized model must be selected from a finite data set, with the consequent problem of over-fitting -the model becoming too strongly tailored to the particularities of the training set and generalizing poorly to new data. The SVM principle addresses this problem by balancing the model's complexity against its success at fitting the training data" [24]. SVM does not seek to minimize the training error, but instead it tries to maximize the margin between the training data and the separating hyper-plane. The dimensionality issue is solved by using nonlinear kernel functions. To achieve a high generalization capacity by using optimal separating hyper-planes, the space of input examples is mapped to a space with higher dimensions. If an appropriate mapping is chosen, the high-dimensional space presents the linearly or almost linearly separable input examples. Consequently, the SVM learning is transferred into a quadratic optimization problem having linear constraints with only one global solution. SVM has been used as a universal approximator for various kernels. A subset of the learning data, called support vector, defines SVM and the absence of a local minima is one of its main features. The training data represents the SVM model sparsely and a condensed dataset is extracted from it based on the support vectors [22]. SVM, which was originally suggested for binary classification problems, seeks to find the optimal hyper-plane that defines the borders or the margin between two example classes. When two classes are linearly separable optimal separating hyper-planes can easily define the borders of the classes. However, in some cases the classes may overlap and to overcome the problem of non-separable classes the support vector machine technique is used to produce nonlinear boundaries by creating a linearly separable boundary in a transformed feature space [23]. In this study, WEKA was used as the medium of computation and SVM-SMO (Sequential Minimal Optimization), which is a built-in function of WEKA, is the fastest for the linear SVMs and sparse datasets. SVM evaluation dominates the computation complexity of the SVM-SMO while the required amount of memory for SVM-SMO is linear in the size of the training set allowing it to handle very large training sets such as financial time series data used in this study [22]. "The sequential minimal optimization technique implements John Platt's algorithm to train a support vector classifier. The global implementation replaces all missing values and nominal attributes are transformed into binary ones. All attributes are normalized by default and the output coefficients are based on the normalized data but not on the original data which is crucial for interpreting the classifier" [25].

Results and Discussion

The relevance and quality of the data, usually, has a big impact on the performance of the model used. Thus, the choice of data becomes the most important part in forecasting the markets. In this study, besides ten technical market indicators, seven macroeconomic variables, three international market's close price index values and a sliding window of the last ten days of the ISE National 100 close price index is also included in the total of thirty input attributes to test our models. All series are real-valued and the input data spans from 02/01/1997 to 31/12/2007. For WEKA testing, the statistical model adequacy metrics Root average error (RAE), Root relative squared error (RRSE), accuracy (rate of correctly classified instances) and the square of the correlation coefficient (R^2) are utilized, showing the ability of the model to capture the data. Both for classification and regression experiments a dataset of 10, 20 and 30 inputs were tested in order to see which attribute sets had better predictive power over the other sets.

1. Results for Classification

In this study, simple logistic, ANN and SVM classifiers are used for predicting the market movement direction when different input variables such as technical market indicators, a sliding window of last 10 days and some macroeconomic variables (10 variables) are applied. These features are used to produce the total feature set characterizing the stock market. Simple logistic, ANN and SVM classifiers are trained with the expectation of getting more precise

forecasting results in terms of the market movement direction. In order to calculate the performance of our approach, K-fold cross-validation, which is a well-known method for evaluation, is utilized. K-fold cross validation is used by numerous researchers to reduce the bias related with random sampling of the training and test sets. The test performance of the models is determined by the computation of the following statistical parameters: RAE, RRSE and accuracy (rate of correctly classified instances). The forecasting accuracy can be determined by dividing the number of correctly classified data by the number of the total data. The values of the correctly classified instances for MLP are given in Table 2, while the SVM results are given in Table 3. The SVM and ANN techniques show better performance than the Simple Logistic technique (Table 1), as expected. While the Simple Logistic technique presents only 78.2 % classification accuracy, SVM has 84.1 % and MLP has 84 % correctly classified instances in the best cases. For ANN classification, WEKA's Multilayer Perceptron is used with a learning rate of 0.1 and a momentum value of 0.7 with number of the neurons in the hidden layers of 10, 20,.,90. In table 2, the model produces its highest value with a 84 % classification power for 40 neurons in its hidden layer using technical indicators and the last 10 sliding window variables. It also shows that the combination of technical market indicator inputs and the last 10 sliding window inputs provide seemingly better performance (84 %) than [3]'s average BPN value of 75.74 %, where only technical indicators are used as input attributes. Even when only technical indicators are employed, MLP shows better performance (80.9 %) which is also over the above figure. It should be noted that without using technical indicators (Table 2), classification results turn out to be very unsuccessful showing the significance of technical market indicators in forecasting market direction. For SVM classification, WEKA's SMO tool is utilized and results for different C values is obtained with all other WEKA default variables kept unchanged. The relevant results can be seen in Table 3. While for C values up to 50, most results are almost identical for all input combinations except for the feature set where technical market indicators were not utilized, better results are obtained for C values above 100. A peak value of 84.1 % correctly classified instances is found for C=500 in WEKA, that is also superior than [3]'s average SVM value of 71.52 %. The SVM-SMO model is also better when only market indicators were used as inputs presenting 78.9 % success in classification. Checking the results from Table 3, it can be concluded that the macroeconomic variables have no significant effect on the model performance, but rather the sliding window improved the results substantially for C values above 100. The performance demonstrated by these models for forecasting the market movement direction is affected by a couple of factors: input variable choice, forecasting method selection and the best parameter selection. The attributes, which better suit for forecasting the market movement direction, should be used as the inputs of the model. For this reason, along with technical market indicator inputs, the last 10 sliding window inputs and some macroeconomic variables are selected, under the assumption that they are appropriate for forecasting the market movement direction. The advantages of SVM over the simple logistic classifier and ANN methods make it a better tool to map a relationship between the parameters and the features. The combined use of technical market indicators, the last 10 sliding window inputs and several macroeconomic variables with SVM for predicting the market movement direction produces a higher performance of the derived forecasting system. One of the most important properties of SVM is its capability to process high-dimensional data but without dimensionality reduction, which is important in forecasting the market direction. Forecasting can be validated directly by using technical market indicator inputs (10 variables), the last 10 sliding window inputs and possibly a better choice of some macroeconomic variables (here 10 variables are used). The use of technical market indicator inputs (10 variables) improves the performance; also decreases the amount of complexity and simplifies the calculation. Besides, this technique tries to extract the most valuable characteristic input features by minimizing redundancy and exclude noise from the stock market. In general, all techniques accomplish a good performance up to 84.1%. A slightly lower performance is observed when the simple logistic classifier is applied as compared to other data mining tools. Table 1 shows the performances of this estimator using different attributes of the stock market (ISE) as input features. Accurate identification of stock market movement direction is important for both forecasting and evaluation. The forecasting accuracy improves significantly when technical market indicator inputs (10 variables) are used, providing 78.9% accuracy. The effect of the feature selection with technical market indicator inputs (10 variables), the last 10 sliding window inputs and macroeconomic variables (10 variables) can be

seen from Table 3, giving the best results for SVM. The enhanced forecasting accuracy of the SVM using technical market indicator inputs (10 variables) as basic stock market parameters makes it an attractive alternative for forecasting the stock market direction by increasing the effectiveness of the estimation.

Similar studies of linear regression, MLPNN (MLP neural network) and SVM for forecasting the stock market movement direction are available as explained in the literature review. This study is about comparing the stock market direction prediction abilities of some machine learning techniques as well as predicting the stock index price levels of ISE (regression). Improved performance using different machine learning tools also suggests the importance of nonlinear approaches for modeling the relationships between technical market indicators (10 variables), the last 10 sliding window inputs, macroeconomic variables and the ISE stock market characteristics. Based on this study, it is reasonable to conclude that further advances in forecasting stock market direction may be achieved through the incorporation of two approaches. The first is input feature selection for separating relevant features to improve the prediction power of the model. The second is to choose the appropriate forecasting technique for predicting the market movement direction. Considering the results of the present work and similar stock market movement direction forecasting problems, the followings can be emphasized:

1. The high forecasting accuracy of the SVM classifier gives insights into the features used for defining the stock market data. The results drawn in the applications demonstrated that the technical market indicators are the features, which represent the stock market data well, and by the use of these features a good distinction between each direction can be obtained.

2. Simple logistic, ANN and SVM based estimators are appropriate for use in forecasting stock market movement direction; but, SVM has an advantage over other forecasting methods based on its higher forecasting accuracy.

3. Simple logistic is an acceptable forecasting method. But, it does not have a good forecasting accuracy and cannot easily handle nominal data types. SVM is based on preprocessing the data to represent patterns in a high dimension typically much higher than the original feature space. With an appropriate nonlinear mapping to a sufficiently high dimension, data from different categories can always be separated by a hyper-plane. As a result, while the original features bring sufficient information for good forecasting, mapping to a higher dimensional feature space make available better discriminatory evidence that are absent in the original feature space. The problem of training an SVM is to select the nonlinear functions that map the input to a higher dimensional space. Often this choice will be informed by the designer's knowledge of the problem domain. Polynomials, Gaussians or other basis functions might be used in the absence of such information. The dimensionality of the mapped space can be arbitrarily high. For training the SVM, appropriate kernel parameters sigma, and C were selected by using the trial and error method. The optimal sigma, and C values can only be ascertained after trying out different values. In addition, the choice of sigma parameter in the SVM is crucial in order to have a suitably trained SVM. The SVM has to be trained for different kernel parameters until to get the best result.

TABLE I
SIMPLE LOGISTIC CLASSIFICATION RESULTS

Input Feature Set	Correctly class. (%)
technical indicators + macroeconomic variables + last 10	78.2
technical indicators + macroeconomic variables	78.2
technical indicators	78
technical indicators + last 10	78
macroeconomic variables + last 10	52

TABLE II
CORRECTLY CLASSIFIED INSTANCES (%) RESULTS USING MLP

Input Feature Set	# of neurons in the hidden layer (n)						
	10	20	30	40	50	70	90
technical indicators + macroeconomic variables + last 10	81	80.1	80.6	81.1	80.5	80.6	80.5
technical indicators + macroeconomic variables	78.8	77.4	78.6	78.3	79.2	79.1	78.7
technical indicators	80.9	80.9	80.7	80.7	80.2	80.8	80.3
technical indicators + last 10	83.9	83.7	82.9	84	83.4	83.9	83.1
macroeconomic variables + last 10	52	52.2	52.2	53.1	53.2	53.2	53

TABLE III
CORRECTLY CLASSIFIED INSTANCES (%) RESULTS USING SVM

Input Feature Set	C values								
	1	5	20	50	70	100	200	300	500
technical indicators + macroeconomic variables + last 10	78.2	79.4	80.6	78.6	78.7	78.7	82.9	83.2	84.1
technical indicators + macroeconomic variables	77.9	78.4	78.3	78.2	78.3	78.5	78.6	78.5	78.7
technical indicators	77.8	78.3	78.3	78.2	78.2	78.3	78.7	78.7	78.9
technical indicators + last 10	77.9	78.3	78.4	78.8	82	82.1	82.7	83.3	84.1
macroeconomic variables + last 10	52.3	52.3	53.1	53.6	53.4	52.6	52.2	53	48.4

2. Results for Regression

As far as regression results are concerned, a similarity measure called the coefficient of determination or the square of the correlation coefficient (R^2) is added to the table results, which should actually be very close to 1 to show strong correlation or a perfect fit as seen in Tables 4-8. Again, the SVM and ANN techniques mostly outperformed the linear regression method (Table 4) in all categories which is an expected outcome, as well. For MLP regression, Tables 5 and 6 prove the effectiveness of the sliding window when used together with technical indicator inputs creating much lower error values. This result can also be observed from Table 7 and 8, showcasing the SVM regression tests. Comparing MLP regression with SVM regression outcomes it can be seen that the SVM model has better regression forecasting power than the MLP model with 0.29 % RAE and RRSE values found for all input attributes and C value of 300. The best MLP regression results are 0.39 % RAE and 0.47 % RRSE for the technical indicators and sliding window combination input set applied on 4 neurons (n=4) in the hidden layer. For both SVM and MLP regression, it is noted that technical market indicators play an important role in forecasting the price levels. However, macroeconomic data showed no significant improvement in the overall results. The results also indicate that for lower number of neuron values (n) the MLP regression predictive power improves

significantly, as well. As for the figures indicating the real and estimated values of a string of daily close values, SVM again proves itself as a quite precise estimator with almost a perfect fit value (R^2) of 1 while MLP also shows strong mapping ability between the real and estimated values much superior to the linear regression method. A paired t-test was conducted to assess the level of significance regarding the SVM, MLP and Linear Regression performances at Table 9. The hypothesis that the mean accuracy of the SVM is equal to MLP and Linear Regression, has been significantly rejected on a 95% confidence level (α = 0:05) proving the superiority of SVM to the other two methods. In fact, MLP also provides good prediction results but not as good as SVM as seen from the table.

TABLE IV
LINEAR REGRESSION RESULTS

Input Feature Set	RAE (%)	RRSE (%)
technical indicators + macroeconomic variables + last 10	1.9	2.3
technical indicators + macroeconomic variables	2.6	3
technical indicators	2.6	3.1
technical indicators + last 10	1.9	2.3
macroeconomic variables + last 10	2.5	3

TABLE V
MLP REGRESSION RESULTS (% RELATIVE ABSOLUTE ERROR VALUES – % RAE)

Input Feature Set	# of neurons in the hidden layer (n)							
	4	7	10	20	40	50	70	90
technical indicators + macroeconomic variables + last 10	1	0.87	1.06	1.15	1.13	1.24	0.94	1.33
technical indicators + macroeconomic variables	1.80	1.61	1.70	1.76	1.88	1.90	1.78	1.83
technical indicators	1.71	1.63	1.74	2	2.32	2	2	2.1
technical indicators + last 10	0.39	0.42	0.42	0.6	0.73	0.75	1.84	1.63
macroeconomic variables + last 10	3.46	3.35	3.33	3.41	3.55	3.60	3.41	8.9

TABLE VI
MLP REGRESSION RESULTS (% ROOT RELATIVE SQUARED ERROR – %RRSE)

Input Feature Set	# of neurons in the hidden layer (n)							
	4	7	10	20	40	50	70	90
technical indicators + macroeconomic variables + last 10	1.05	0.95	1.20	1.29	1.24	1.35	1.05	0.95
technical indicators + macroeconomic variables	1.73	1.91	1.79	1.87	1.95	1.98	1.73	1.91
technical indicators	1.86	1.80	1.91	2.22	2.46	2.1	1.86	1.80
technical indicators + last 10	0.47	0.49	0.49	0.69	0.83	0.87	0.47	0.49
macroeconomic variables + last 10	3.81	3.70	3.70	3.79	3.96	4	3.81	3.70

TABLE VII
SVM REGRESSION RESULTS (% RELATIVE ABSOLUTE ERROR VALUES – Л % RAE)

Input Feature Set	C values					
	10	20	100	200	300	500
technical indicators + macroeconomic variables + last 10	0.36	0.33	0.31	0.30	0.29	0.30
technical indicators	1.46	1.46	1.46	1.46	1.46	1.46
technical indicators + last 10	0.36	0.33	0.32	0.31	0.31	0.31
macroeconomic variables + last 10	2.45	2.45	2.45	2.45	2.45	2.45

TABLE VIII
SVM REGRESSION RESULTS (% ROOT RELATIVE SQUARED ERROR VALUES - % RRSE)

Input Feature Set	C values					
	10	20	100	200	300	500
technical indicators + macroeconomic variables + last 10	0.4	0.34	0.31	0.29	0.29	0.30
technical indicators	1.9	1.9	1.9	1.9	1.9	1.9
technical indicators + last 10	0.40	0.34	0.31	0.30	0.30	0.30
macroeconomic variables + last 10	3.1	3.1	3.1	3.1	3.1	

TABLE IX
STATISTICAL SIGNIFICANCES FOR 30 FEATURES

Best Statistical Significance	SVM	MLP	Linear Regression
RAE (%)	0.29	1.0	1.9
RRSE (%)	0.29	1.05	2.3
Paired t-test t-statistics	-7.362	2.556	-0.01144
Paired t-test p-value (two-tailed)	0	0.0106	0.9908

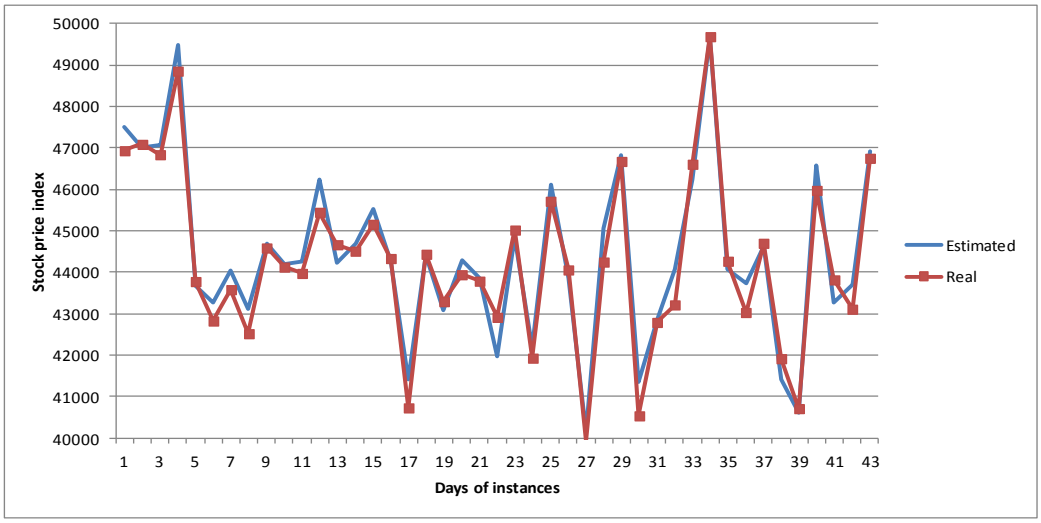

FIGURE 1. Linear Regression result for 30 features
(technical indicators + macroeconomic variables + last 10 sliding window)

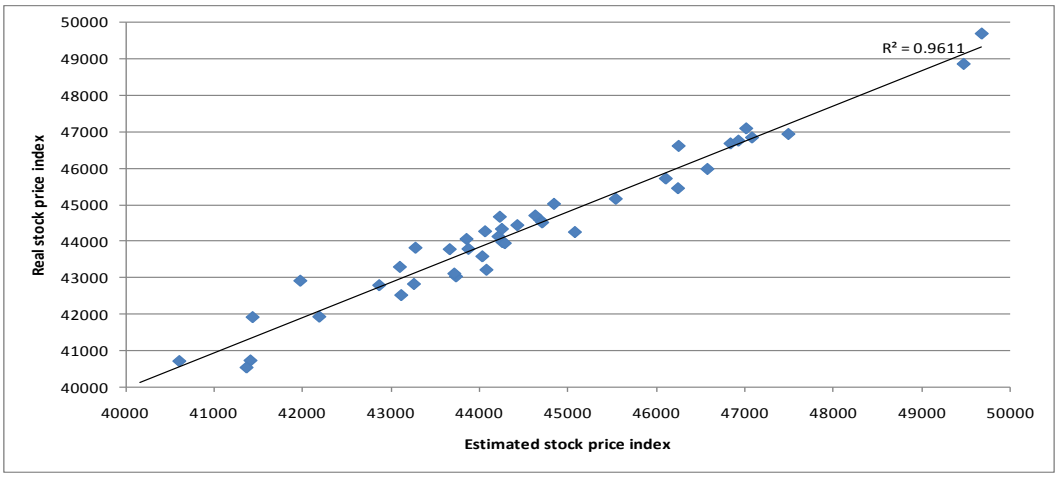

FIGURE 2. Linear Regression R^2 result for 30 features
(technical indicators + macroeconomic variables + last 10 sliding window)

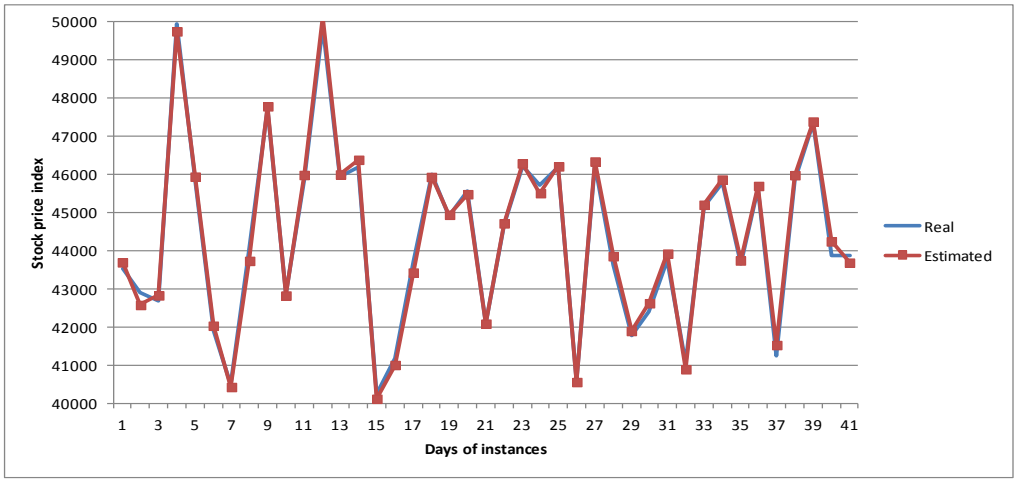

FIGURE 2. MLP Regression result for n=4 (4 neurons in the hidden layer) and 30 features (technical indicators + macroeconomic variables + last 10 sliding window)

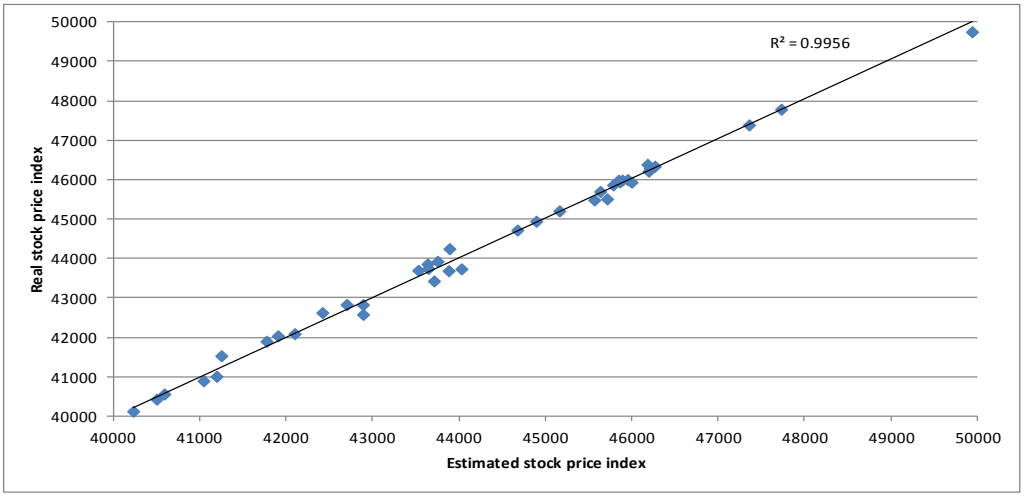

FIGURE 3. MLP Regression R^2 result for n=4 (4 neurons in the hidden layer) and 30 features (technical indicators + macroeconomic variables + last 10 sliding window)

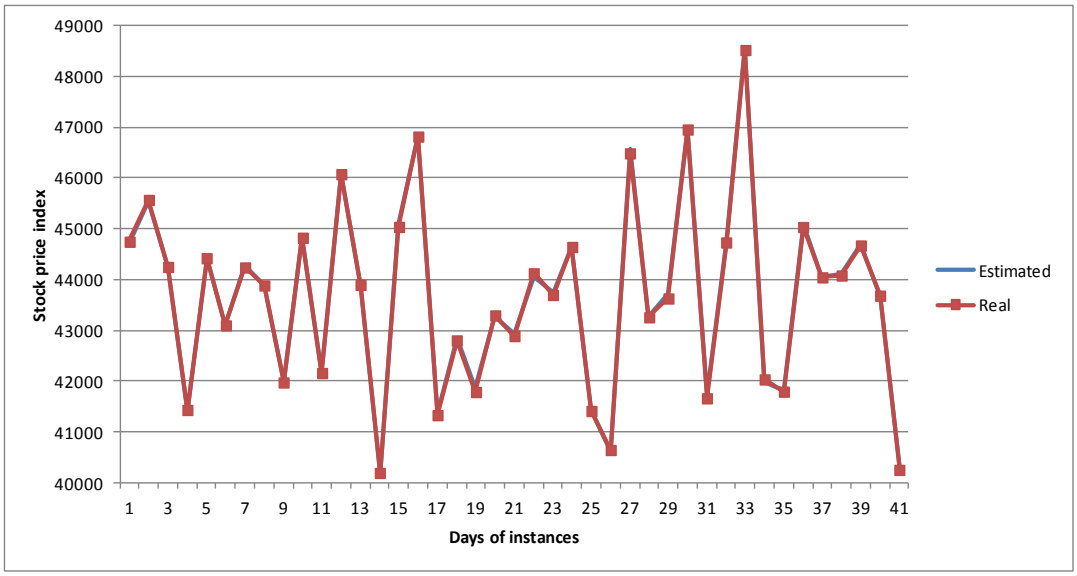

FIGURE 4. SVM for C=300 and 30 features
(technical indicators + macroeconomic variables + last 10 sliding window)

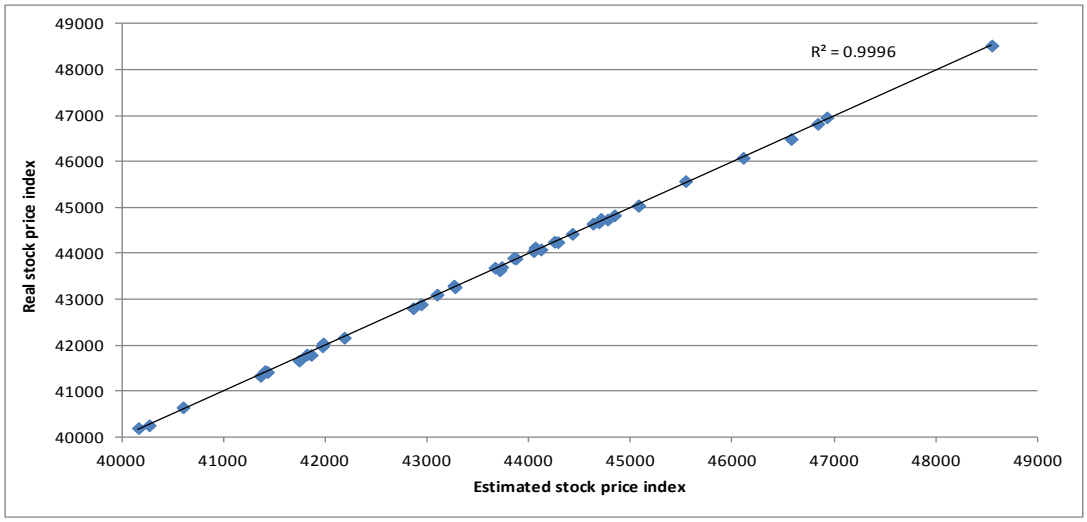

FIGURE 5. SVM R^2 result for C=300 and 30 features
(technical indicators + macroeconomic variables + last 10 sliding window)

Conclusion

The issue of accurately predicting the stock market movement directions is highly important for formulating the best market trading solutions. It is fundamentally affecting buy and sell decisions of an instrument that can be lucrative for investors. Another aspect of this task is to reduce the risk factor involved inherent to the markets. The related study of estimating financial time series data is usually chaotic and complex. This study focused on predicting the ISE National 100 close index direction and closing price levels using classification and regression techniques based on the daily data from 1997 to 2007. The experimental results give us some very important clues. Firstly, both the ANN and SVM models showed superior predicting power in forecasting the stock market movement direction and the stock market price level index, though SVM presented better classification and regression results over MLP. The best values for classification were found to be 84 % both for the SVM and MLP models that is a significant improvement over [3]'s average results of 71.52 % for SVM and 75.74 % for BPN. In case of regression, SVM resulted in 0.29 %

RAE, while MLP presented 0.39 % RAE in the best cases, which are perfectly good outcomes. The t-test result shows the superiority of SVM to the other two methods. In fact, MLP also provides good prediction results but not as well as SVM. Even though the prediction performance of the SVM and ANN models used in this study outperforms studies alike in literature, it is still likely that the forecasting performance of the models can be improved by the following tasks. Either the model parameters should be adjusted by thorough experimentation or the input variable sets need to be modified by selecting those input attributes that are more realistic in reflecting the market workings. [3] had already proved the significance of using ten particular technical market indicators which gave also good results in this study, as well. Besides, the use of a sliding window of the last ten elements of the ISE 100 index proved to be an effective tool in forecasting the market level and direction. However, the seven macroeconomic variables and three other international market indices were not found to be very useful in this study, which means that more appropriate variables have to be found that may improve the forecasting performance of the models employed that can be a further subject of study for interested readers. This study also depicts the reality that simpler methods such as linear regression and Simple Logistic classification becomes inferior to the SVM and ANN structure.

Acknowledgment

We sincerely deliver our special thanks to Dr. Melek Acar Boyacioglu for sharing her data with us.

References:

1. Abu-Mostafa Y. S. and Atiya A. F. (1996), Introduction to financial forecasting, Applied Intelligence, 6(3), 205–213.

2. Tan T. Z., Quek C., and See Ng. G. (2007), Biological brain-inspired genetic complementary learning for stock market and bank failure prediction, Computational Intelligence, 23(2), 236–261.

3. Kara Y., Boyacioglu M.A., and Baykan O.K. (2010), Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange, Expert Systems with Applications 38, 5311–5319.

4. Chen A. S., Leung M.T., and Daouk H. (2003), Application of neural networks to an emerging financial market: Forecasting and trading the Taiwan Stock Index, Computers & Operations Research, 30(6), 901–923.

5. Nyberg H. (2008), Forecasting the direction of the US stock market with dynamic binary probit models, International Journal of Forecasting 27, 561–578.

6. Leung M. T., Daouk H., and Chen A. S. (2000), Forecasting stock indices: A comparison of classification and level estimation models, International Journal of Forecasting, 16, 173–190.

7. Hong Y. and Chung J. (2003), Are the directions of stock price changes predictable? Statistical theory and evidence, Cornell University, Unpublished manuscript.

8. Rydberg T. and Shephard N. (2003), Dynamics of trade-by-trade price movements: decomposition and models, Journal of Financial Econometrics, 1, 2–25.

9. Kauppi H. and Saikkonen P. (2007), Predicting US recessions with dynamic binary response models, Review of Economics and Statistics, 90, 777–791.

10. Altay E. and Satman M. H. (2005), Stock market forecasting: Artificial neural networks and linear regression comparison in an emerging market, Journal of Financial Management and Analysis, 18(2), 18–33.

11. Avci E. (2007), Forecasting daily and sessional returns of the ISE-100 index with neural network models, Journal of Dogus University, 8(2), 128–142.

12. Karaatli M., Gungor I., Demir Y., and Kalayci S. (2005), Estimating stock market movements with neural network approach, Journal of Balikesir University, 2(1), 22–48.

13. Diler A.I. (2003), Predicting direction of ISE national-100 index with back propagation trained neural network, Journal of Istanbul Stock Exchange, 7(25–26), 65–81.

14. Cao Q., Leggio K.B., and Schniederjans M. J. A. (2005), A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market, Computers and Operations Research, 32, 2499–2512.

15. Kim K. (2003), Financial time series forecasting using support vector machines, Neurocomputing, 55, 307–319.

16. Manish K. and Thenmozhi M. (2005), Forecasting stock index movement: A comparison of support vector machines and random forest, in Proceedings of ninth Indian institute of capital markets conference, Mumbai, India.

17. Huang W., Nakamori Y. and Wang S. Y. (2005), Forecasting stock market movement direction with support vector machine, Computers and Operations Research, 32, 2513–2522.

18. Hsu S. H., Hsieh J. J. P. A., Chih T. C., and Hsu K. C. (2009), A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression, Expert Systems with Applications, 36(4), 7947–7951.

19. Boyacioglu M.A. and Avci D. (2010), An Adaptive Network-Based Fuzzy Inference System (ANFIS) for the prediction of stock market return: The case of the Istanbul Stock Exchange, Expert Systems with Applications 37, 7908–7912.

20. Yumlu, S., Gurgen F. and Okay N. (2005), A comparison of global, recurrent and smoothed-piecewise neural models for Istanbul stock exchange (ISE) prediction, Pattern Recognition Letters 26, 2093–2103.

21. McNelis P. D. (2005), Neural Networks in Finance-Gaining Predictive Edge, Elsevier Academic Press, 13-15.

22. Du K. L. and Swamy M. N. S. (2006), Neural Networks in a Softcomputing Framework, Springer-Verlag.

23. Hastie T., Tibshirani R., and Friedman J. (2008), The Elements of Statistical Learning, Springer, 2nd. Edition.

24. Vapnik and Chervonenkis A. (1974), Theory of Pattern Recognition, Nauka, Moscow.

25. Weka (2010), Waikato Environment for Knowledge Analysis, Version 3.7.3, The University of Waikato Hamilton, New Zealand.