Applied Soft Computing 67 (2018) 337–349

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A hybrid financial trading support system using multi-category classifiers and random forest

Manoj Thakur*, Deepak Kumar*

Indian Institute of Technology-Mandi, Mandi 175001, Himachal Pradesh, India

ARTICLE INFO

Article history:
Received 2 August 2017
Received in revised form
17 December 2017
Accepted 3 March 2018
Available online 10 March 2018

Keywords:
Financial forecasting
Technical Analysis
Weighted Multicategory GEPSVM
Random forest
Walk forward approach

ABSTRACT

This study presents a decision support system for algorithmic trading in the financial market that uses a new hybrid approach for making automatic trading decision. The hybrid approach integrates weighted multicategory generalized eigenvalue support vector machine (WMGEPSVM) and random forest (RF) algorithms (named RF-WMGEPSVM) to generate "Buy/Hold/Sell" signals. The WMGEPSVM technique has an advantage of handling the unbalanced data set effectively. The input variables are generated from a number of technical indicators and oscillators that are widely used in industry by professional financial experts. Selection of relevant input variables can enhance the predictive capability of the prediction algorithms. RF technique is employed to discover the optimal feature subset from a large set of technical indicators. The proposed hybrid system is tested using "walk forward" approach for its capability of taking an automatic trading decision on daily data of five index futures, viz., NASDAQ, DOW JONES, S&P 500, NIFTY 50 and NIFTY BANK. RF-WMGEPSVM achieves the notable improvement over the buy/hold strategy and other predictive models contemplated in this study. It is also observed that combining WMGEPSVM with RF further improves the results. Empirical results confirm the effectiveness of RF-WMGEPSVM in the real market scenarios having bullish, bearish or flat trend.

1. Introduction

Over the last few decades, the growing popularity of computing technologies and electronic communication networks has made buying or selling of financial assets much easier and faster. This had led to ample trading opportunities to both individual investors and financial institutions, and now a days requires nothing more sophisticated tools than a computer machine. Due to such consequence, financial securities industries have been among the fastest emerging fields in the modern economic world. The maturity of financial markets is resulting in the shaping of a more complex, nonlinear, noisy, dynamic and chaotic system [1–3]. Hence, the problem of financial market trend prediction has been a difficult job and demands new models that can capture the behavior of securities industry trends and can help to make better trading decisions. Numerous studies have been conducted to model the stock market behavior using machine learning algorithms that can learn useful models from price data. Research done in this area has led to sophisticated techniques with reduced computation time and improved the capability of systems to address the intricate problem offered in stock markets to trend prediction.

One of the most widely used learning algorithm for forecasting the trend of financial market is artificial neural networks (ANN) [4–7]. ANN has the capability of approximating any non-linear function without any prior assumptions about the model, underlying function, and the input data and thus many researchers [8–11] have applied ANNs to predict the future market trend or price. support vector machine (SVM) [12], a structural risk minimization (SRM) principle [13] based machine learning algorithm has recently gained popularity in financial modeling than ANNs [14–16,5] in the last decade. SVM has been reported to have better performance than ANNs in financial forecasting [17–19,16] due to low generalization error as compared to empirical risk minimization techniques. Deriving features from quantitative information available in the market such as historical prices have shown promising results. Several studies [20,17,18,21,19,22,23] have demonstrated that implementation of SVM to forecast trends or the prices of the assets by taking technical indicators as input variables outperform the ANNs.

Deciding the appropriate feature space for SVM [16,5,15] as an input for financial decision support system is critical due to its sensitive correlation among input variables and inherent noise of the financial time series. The low correlation of input to output

* Corresponding authors.
E-mail addresses: manojpma@gmail.com (M. Thakur), deepaktyagi12@gmail.com (D. Kumar).

https://doi.org/10.1016/j.asoc.2018.03.006
1568-4946/

or high correlation with another input may drastically affect the performance of the prediction model and may result in high generalization error in the prediction phase. Feature selection techniques try to address this problem by discovering the optimal subset of input variables. Therefore, research in the recent years is also focused on the development and use of suitable feature selection techniques to enhance the generalization capability and reducing the computation time in building the prediction model for financial markets. Various feature extraction methods such as Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA), Independent Component Analysis (ICA) [24], Genetic Algorithm (GA) [14] and nonlinear independent component analysis [25], have been used to identify the optimal subset of features that serve as inputs to SVMs for forecasting the market prices. Recently, Lee [26] proposed SVM based prediction model with hybrid feature selection techniques for forecasting the future direction that combines the *F*-score with Supported Sequential Forward Search (F-SSFS).

In the last decade, research in the area of automatic learning has opened an unprecedented possibility in developing more reliable decision support system for financial trading strategy. Various hybrid artificial intelligence tools have been developed that help to make a financial trading decision. Tsaih et al. [27] applied a hybrid AI approach that integrates the neural network technique with the rule-based systems technique to predict daily price changes in S&P 500 index futures. A decision support system was proposed by Baba et al. [28], combining neural networks and genetic algorithms for forecasting the Tokyo Stock Exchange Indexes. Though many studies for trading decision support system have shown encouraging results, most of them are evaluated based on statistical measures such as forecasting accuracy, mean squared error, precision, recall. Transaction cost is one of the most important factors that needs to be considered while evaluating the performance of a financial trading system. Most of the studies done in the area of automatic trading system have not included transaction cost while measuring the trading system performance. However, Alexander [29] and Hudson et al. [30] are reported that the performance of a trading strategy is quite sensitive to the transaction cost.

In more recent work, a hybrid neurogenetic system for financial trading was proposed by Kwon et al. [31]. A recurrent neural network (RNN) is used for the predicting the market direction and a Genetic Algorithm (GA) optimizes the NN's weights under a 2-D encoding and crossover. Wen et al. [32] proposed an intelligent stock decision support system by combining the support vector regression (SVR) [33] and box theory. Both studies have been reported that automatic stock decision support system is performed significantly better than buy-and-hold strategy. Above mentioned studies have concluded that the hybrid prediction model in financial forecasting performs better than the individual learning paradigm. However, the way of integration of such methods is very critical for development of such prediction models. The aim of these studies is to combine use of different techniques to achieve better decision making system than employment of the techniques alone. Recently, Kumar et al. [17] proposed Proximal support vector machine (PSVM) based predictive model with random forest that uses a tree based ensemble technique to rank the relative importance of input technical indicators. The empirical findings suggested that random forest (RF) in combination with PSVM is superior to linear correlation, rank correlation, regression relief techniques for stock index trend prediction.

This study focuses on the development of financial trading systems which uses Weighted Multicategory GEPSVM (WMGEPSVM) [34] in combination of random forest (RF) and technical indicators. This automatic financial trading system is named as RF-WMGEPSVM. Various technical indicators and oscillators are used as input variables for WMGEPSVM to predict the trends of the market. The RF technique is applied to find the optimal set of input technical indicators for WMGEPSVM. The trading strategy gives "Buy/Hold/Sell" signals that are generated based on the historical price of the stock indices. The proposed financial trading system is tested with the NASDAQ, DOW JONES and S&P 500, NIFTY 50 and NIFTY BANK futures indices over a period of 500 trading days. The data considered for the study consist of variety of market scenarios that are witnessed during various phases (bullish, bearish and sideways) of the real financial markets. Some of the most widely used practical performance measures used in financial markets such as rate of return, percent profitability and Maximum drawdown are used to evaluate the performance of the proposed trading strategy. Performance of RF-WMGEPSVM is compared with traditional and recently developed approaches such as "buy and hold" investment strategy, RF-PSVM [17], balanced multicategory support vector machine (BMPSVM) [35], OVA-Multi-class least squares twin SVM (MLSTSVM) [36] and RF in combination with BMPSVM (RF-BMPSVM) and MLSTSVM (RF-MLSTSVM) based on these measures. Experimental findings show that the RF-WMGEPSVM system outperforms most of other strategies with respect to these matrices and is also found to be effective in bullish, bearish and sideways financial market scenarios.

The contents of this paper are organized as follows. The proposed financial trading strategy are explained in Section 2. Section 3 gives a brief introduction of RF and WMGEPSVM used in the current study. Section 4 reports the experimental findings comparison and analysis of results. Conclusions from the current study are drawn in Section 5.

2. Financial trading decision system using RF-WMGEPSVM

The decision making in financial markets involves picking the right stock for possibility of getting desired returns. There are two different type of techniques used for making an investment decision. These methodologies are termed as Fundamental Analysis and Technical Analysis [37].

Fundamental Analysis is a way of analyzing the price of the stocks based on various economical factors such as cash flow, balance sheet and income analysis. In Fundamental Analysis the reports about the profits, expenditures, holding of resources having a potential of future earnings, debts and interests paid against a debt are taken in to account to take an investment decision.

Technical Analysis is based on postulates collected from the Dow Theory given by Charles Dow around a hundred years back (refer [38] for details). Technical Analysis deals with the study of price movements in the market. It only takes into account the chart patterns generated with the historical stock price data. On a chart, the price is plotted in the form of a bar that enables the analysis of a chart on a fixed time frame. A green bar is called a bullish bar (indicates a rise in the price) indicating that the close price on the bar is higher than the open price. On the other hand a red bar is called a bearish bar (points to a fall in the price) reflecting the scenario when the close price is less than the open price. The peak and bottom of the chart reflect the high and low prices seen during the period of the bar. Fig. 1 reflects the bullish, bearish and sideways market trend for Dow Jones index future based on chart made up of daily data.

A trending market is one which moves in only one direction. A bullish (up) trend is identified by a series of higher high and higher low while a bearish (down) trend is identified by a series of lower high and lower lows. When the stock price moves in a narrow range (a range bound market), the price trend is difficult to identify. Another major difficulty faced is to differentiate between correction and trend reversal. A correction is a movement in a trending market when the price pulls back a little and resumes moving again

Fig. 1. Chart of Dow Jones index future based on daily data.

in that direction. On the other hand, a reversal is a price moment when stock price start moving the opposite trend (uptrend in case of bearish market and down trend in case of bullish market). The chart pattern and trends formed are analyzed to predict the price movement.

The chart is analyzed using a set of instruments termed as technical indicators that are superimposed onto a chart to spot identifiable price patterns. Based on the collective behavior of the price data and technical indicators trading strategies are made. A trading strategy is aimed to capture the major trends of the market. Technical Analysis is about the decision making based on price trend. The objective is to make a decision (a buy signal or a sell signal) based on the price trend, and then keep the position as it is until it becomes evident from a weight of the evidences suggested by a number of technical indicators indicating the change in the trend (i.e. the current trend has ended and a reverse trend has begun). For comprehensive details about the Technical Analysis, readers are advised to refer [37,39,40].

Financial institutions and investors require to make investment decisions depending on their behavioral scopes and trading horizon such as intraday, daily, weekly, and monthly trading. In this study, a trading strategy is considered which make use of daily data index future.

The daily market data at day t consists of open (O_t), high (H_t), low (L_t), close (C_t) price and volume (V_t). The signals at day t, i.e., S_t is produced based on the following rule.

$$S_t = \begin{cases} BUY, & \text{if } C_{(t+1)} > H_{(t)} \text{ and } L_{(t+1)} > L_{(t)} \\ SELL, & \text{if } C_{(t+1)} < L_{(t)} \text{ and } H_{(t+1)} < H_{(t)} \\ HOLD, & \text{otherwise} \end{cases} \quad (1)$$

A schematic representation of a desired trading strategy is depicted in Fig. 1.

2.1. Financial trading strategy

The financial trading strategy, based on the signal S_t, first square off the position (if it has already an opened position) and then take a new position, is depicted in Algorithm 1. If S_t signals BUY, the short position is squared off, if any, and then long position is taken. If S_t signals SELL, long position (if any) is squared off and a short position is taken. If S_t signals Hold then long or short position already taken is kept as it is till these signal of square off (buying in case of short positions and selling in the case of long position).

Algorithm 1. "BUY-HOLD-SELL" trading strategy

```
Initialize LONG ⟵ NONE
    SHORT ⟵ NONE
    BUY ⟵ NONE
    SELL ⟵ NONE
IF (S_t = Buy)
  IF(SHORT ≠ NONE)
    SELL = C_t      // Short position is squared off at price C_t.
    SHORT ⟵ NONE
  END IF
  IF(LONG = NONE)
    BUY = C_t       // Long position is opened at price C_t.
    LONG ⟵ 1
  END IF
END IF
IF (S_t = Sell)
  IF(LONG ≠ NONE)
    LONG = C_t      // Long position is squared off at price C_t.
    LONG ⟵ NONE
  END IF
  IF(SHORT = NONE)
    SELL = C_t      // Short position is opened at price C_t.
    SHORT ⟵ 1
  END IF
END IF
IF (S_t = Hold)
  SHORT
  LONG        // Keep all positions open.
END IF
```

The performance of a financial trading system is evaluated based on some of the most significant performance measures such as the rate of return (ROR), maximum draw down (MDD) and percent profitability (PP). Let R_0 is the initial investment and R_t be the value of fund at day t.

The ROR is one of the most useful criterion for evaluating a trading strategy that indicates the profit/loss on the initial investment over a period of time. It can be calculated for a period of time t.

$$ROR = \frac{R_t - R_0}{R_0} * 100 \quad (2)$$

Maximum drawdown (MDD) is an important criterion to evaluate the trading system that is often used as downside risk measure over a specified time period t. MDD is expressed as the maximum

Table 1
Description and input variable formulation.

S. no.	Input variable formulation	Description
1	$x_{(t,1)}$ = open price at time t	Open price
2	$x_{(t,2)}$ = high price at time t	High price
3	$x_{(t,3)}$ = low price at time t	Low price
4	$x_{(t,4)}$ = closing price at time t	Closing price
5	$x_{(t,5)}$ = volume at time t	Trade volume
6	$x_{(t,6)} = \frac{1}{\delta}\sum_{i=1}^{\delta} CL_{t-i}$	δ-days moving average (SMA)
7	$x_{(t,7)} = \alpha(CL_t - EMA_{t-1}) + EMA_{t-1}$	δ-days exponential moving average (EMA) where $\alpha = \frac{2}{\delta+1}$
8	$x_{(t,8)} = \frac{CL_t - CL_{t-y}}{CL_{t-\delta}} \times 100$	δ-days Relative Difference in percentage (RDP)
9	$x_{(t,9)} = \frac{CL_t - LL_{t-\delta}}{HH_{t-\delta} - LL_{t-\delta}}$	Stochastic $\%K$
10	$x_{(t,10)} = \frac{\sum_{i=0}^{i=\delta-1} \%K_{t-i}}{\delta}$	The moving average of $\%K$ ($\%D$)
11	$x_{(t,11)} = \frac{HH_{t-\delta} - CL_t}{HI_{t-\delta} - LL_{t-\delta}}$	Larry William's $\%R$
12	$x_{(t,12)} = \frac{CL_t - MA(\delta)}{y}$	δ-days Bias
13	$x_{(t,13)} = (EMA(\delta) - EMA(\delta - d))$	δ days Moving Average Convergence and Divergence $MACD$
14	$x_{(t,14)} = CL_t - CL_{t-\delta}$	Momentum measures change in stock price over last δ days (MTM)
15	$x_{(t,15)} = \frac{CL_t}{CLt-\delta} \times 100$	δ-days Price Rate of Change
16	$x_{(t,16)} = \frac{MA(\delta) - MA(z)}{MA(\delta)}$	Price oscillator ($OSCP$)
17	$x_{(t,17)} = \frac{1}{2}(LL_{(t-\delta)} + HH_{(t-\delta)})$	Median Price (MP)
18	$x_{(t,18)} = min(LO_{(t-\delta)})$	Lowest price (LL) of last δ-days
19	$x_{(t,19)} = max(HI_{(t-\delta)})$	Highest price (HH) of last δ-days
20	$x_{(t,20)} = \frac{CL_t - MA(\delta)}{0.015\sigma}$	Commodity channel index (CCI)
21	$x_{(t,21)} = \frac{1}{10*SMA(z)}(SMA(\delta) - SMA(z)) + SMA(z)$	A signal line is also known as a trigger line ($signalLine(\delta, z)$)
22	$x_{(t,22)} = \frac{1}{\delta}(ATR_{(t-1)} \times (\delta - 1) + TR_t)$	δ-days average true range (ATR) where $TR_t = max[(H_t - L_t), abs(H_t - C_{(t-1)}), abs(L_t - C_{(t-1)}]$
23	$x_{(t,23)} = 100 \times \frac{(EMA(EMA(MTM(1),\delta),\delta)}{(EMA(EMA(\lvert MTM(1)\rvert,\delta),\delta)}$	True strength index (TSI)
24	$x_{(t,24)} = 100 \times \frac{1}{4+2+1}(4 * avg(\delta) + 2 * avg(\delta) + avg(z))$	Ultimate oscillator (UO) where $avg(t) = \frac{bp_1 + bp_2 + \cdots + bp_t}{tr_1 + tr_2 + \cdots + tr_t}$, $bp_t = CL_t - min(LO_t - CL_{(t-1)})$ $tr_t = max(H_t, CL_{(t-1)} - min(LO_t - CL_{(t-1)})$,
25	$x_{(t,25)} = \sqrt{(R_1^1 + R_2^2 + \ldots + R_n^2)}$ where $R_t(\delta) = \frac{100}{HH_{(t-\delta)}} \times (CL_t - HH_{(t-\delta)})$	δ-days Ulcer index (Ulcer)

$LL_{t-\delta}$ and $HH_{t-\delta}$ are the lowest low and highest high in last δ trading days respectively. $(t-\delta)$ denotes the last δ trading days from day t and different input features can be generated by using the different values of input parameter δ.

cumulative loss from the peak to the subsequent trough. MDD is computed in percentage terms as:

$$PP = \frac{SUP_{s\in[0,t]}X(s) - X(t)}{SUP_{s\in[0,t]}X(s)} * 100 \tag{3}$$

The PP (also called as the probability of winning) is obtained by dividing the total number of winning trades by the total number of trades for the entire trading period t.

$$PP = \frac{\text{Number of winning trades}}{\text{Total number of trades}} \tag{4}$$

3. Methodology

3.1. Random forest

Random forests (RF) [41] is an efficient and extensively used statistical learning algorithm for both categorization and regression problems. RF has been also getting attention in feature pruning due to its better generalization capability and robustness. It constructs an ensemble model of many unpruned decision trees η from random subsets of features, where each tree Γ_i, ($\forall i = 1, 2, \ldots, \eta$) is constructed using bootstrapped training data. In RF, a subset of randomly selected features is taken rather than defining the best split among all the input variables.

Let $((\chi_1, y_1), (\chi_2, y_2), \ldots, (\chi_m, y_m))$ denotes the training set consisting of feature vectors, $\chi \in \mathbb{R}^n$, and output, $y_i \in \mathbb{R}$ ($\forall$, $y_i \in \{1, 2, \ldots, K\}$ into K classes), where m is the number of trading days and n is the number of features. Firstly, data is sampled with replacement for each decision tree Γ_i. For constructing the decision tree, best split among $mtry \ll n$ randomly selected variables one of n features is chosen. The final classification decision is obtained by aggregating the outcomes of all η decision trees. In RF, randomness in the construction of different trees and selecting the best split feature subset plays a critical role for better generalization. Hence, only two-third of randomly split bootstrapped data is utilized for training and remaining one-third data, called Out-of-a Bag (OOB) samples, is utilized to test the performance of a decision tree and is also used to determine the importance score of the variable. The prediction error rate ϱ_{Γ_i} is computed on OOB samples for all decision trees Γ_i ($\forall i = 1, 2, \ldots, \eta$). Then, the OOB samples are perturbed

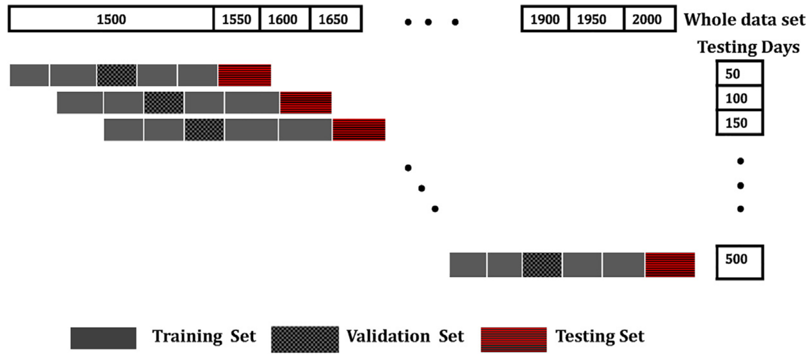


Fig. 2. Model selection and testing scheme used in this study.

for each decision tree Γ_i to obtain the importance score of each feature f_i. Again, the features are randomly permuted among the OOB samples and prediction error ϱ'_{Γ_i} for the perturbed OOB samples are recorded for each decision tree Γ_i. The importance score of a feature IS_f^{RF} is given as follows:

$$IS_f^{RF} = \frac{1}{\eta}\sum_{\Gamma_i}(\varrho_{\Gamma_i} - \varrho'_{\Gamma_i}) \tag{5}$$

The features are sorted according to IS_f^{RF} in descending order. Take down that true features are significantly more informative than the noisy ones.

3.2. Weighted Multicategory GEPSVM

Weighted multicategory generalized eigenvalue proximal support vector machine (WMGEPSVM) [34] is a multi-class classifier which generates K nonparallel separating hyperplanes by using "One-Versus-All" (OVA) approach for all classes. Each plane is closer to its respective class and farthest from the other classes. To minimize the effect of unbalanced classes, WMGEPSVM uses the weight factor which is determined by using a modified balancing technique proposed in [34]. The data samples belonging to ith class χ_i are labeled as +1, and those from the rest of the class $\chi_j \neq \chi_i$ are labeled as −1. The goal of WMGEPSVM is to determine the following non-parallel planes:

$$P_i = \omega_i^T \chi + b_i, \quad i = 1, 2, \ldots, K \tag{6}$$

where $b_i \in \mathbb{R}$ and $\omega_i \in \mathbb{R}^n$. The plane P_i is constructed by solving a quotient minimizing problem having numerator as the squared sum of 2-norm distances of the data points in the ith class and denominator as the sum of the square of two norm distances of the data points in the rest of the classes to the p_i plane. P_i is determined by solving the following minimization problem:

$$\underset{(\omega_i, b_i)\neq 0}{minimize} \frac{||\chi_i\omega_i + e_i b_i||^2/||[\omega_i, b_i]^T||^2}{\sum_{j=1}^K C_j||\chi_j\omega_i + e_j b_i||^2/||[\omega_i, b_i]^T||^2}, \quad i \neq j \tag{7}$$

and

$$C_j = \frac{1}{m_j}, \quad j \in \{1, 2, \ldots, K\} \tag{8}$$

where C_j are the weighing factors used to normalize the sum of squared of 2-norm distances between every data point of jth class with the ith plane P_i and m_j is the number of data points in the jth class. The regularized formulation of the problem (7) using a Tikhonov regularization term $(\nu||[\omega_i, b_i]^T||^2)$ is given as:

$$\underset{(\omega_i, b_i)\neq 0}{minimize} \frac{||\omega_i\chi_i + e_i b_i||^2 + \nu||[\omega_i, b_i]^T||^2}{\sum_{j=1}^K C_j||\chi_j\omega_i + e_j b_i||^2}, \quad i \neq j \tag{9}$$

where ν is used as a regularization parameter such that $\nu \geq 0$.

Let,

$$\alpha_i = [\ \chi_i \quad e_i\ ]^T[\ \chi_i \quad e_i\ ] + \nu I, \quad \beta_j = \sum_{j=1}^K C_j[\ \chi_j \quad e_j\ ]^T[\ \chi_j \quad e_j\ ],$$

and $z_i = \begin{bmatrix} \omega_i \\ b_i \end{bmatrix}$, the solution become:

$$\underset{z_i \neq 0}{minimize} \frac{z_i^T\alpha_i z_i}{z_i^T\beta_j z_i}, \tag{10}$$

where $\alpha_i, \beta_j \in R^{(n+1)\times(n+1)}$. The optimization problem (10) takes the form similar to a Rayleigh quotient problem and can be solved easily by solving following generalized eigenvalue problem:

$$\alpha_i z_i = \lambda_i \beta_j z_i \tag{11}$$

Here, λ_i is the smallest eigenvalue of (11) and z_i is the corresponding eigenvector. Plane P_i is determined by setting the optimal $[\omega_i, b_i]^T = z_i$ in right hand side of Eq. (6) for ith class. By following a similar procedure, all K planes (one plane for each class) can be generated. After getting (ω_i, b_i), a new sample $\chi \in R^n$ can be classified and assigned to class i having the closest distance to the corresponding plane P_i of (6), i.e.,

$$P(\chi) = arg \min_{i=1,2,\ldots,K} \frac{|(\chi^T\omega_i + b_i)|}{||\omega_i||}. \tag{12}$$

3.3. Input data

Selection of the input variables is a very important task in developing a financial trading support system. There are a number of technical indicators and oscillators that can be utilized as input variables for a financial trading system. It is also observed by the practitioners that a certain sets of indicators are more suitable and effective in capturing certain market movements, e.g., moving averages (MA, EMA) and Relative Difference in percentage (RDP) are preferable for trending market (bullish or bearish) whereas Stochastic %K, %D (moving average of %K), BAIS, commodity channel index (CCI), true strength index (TSI) and price oscillator ($OSCP$) oscillators are more effective for extracting price information in sideways trending market scenarios. Therefore, to give the algorithms an exhaustive set of indicators out of which the relevant one may be selected, we utilized fifty five technical indicators extensively used in previous studies [34,16,42] and on the basis of their application in Technical Analysis. These indicators and oscillators are derived from the financial time series data (O_t, H_t, L_t, C_t and V_t), are described in Table 1.

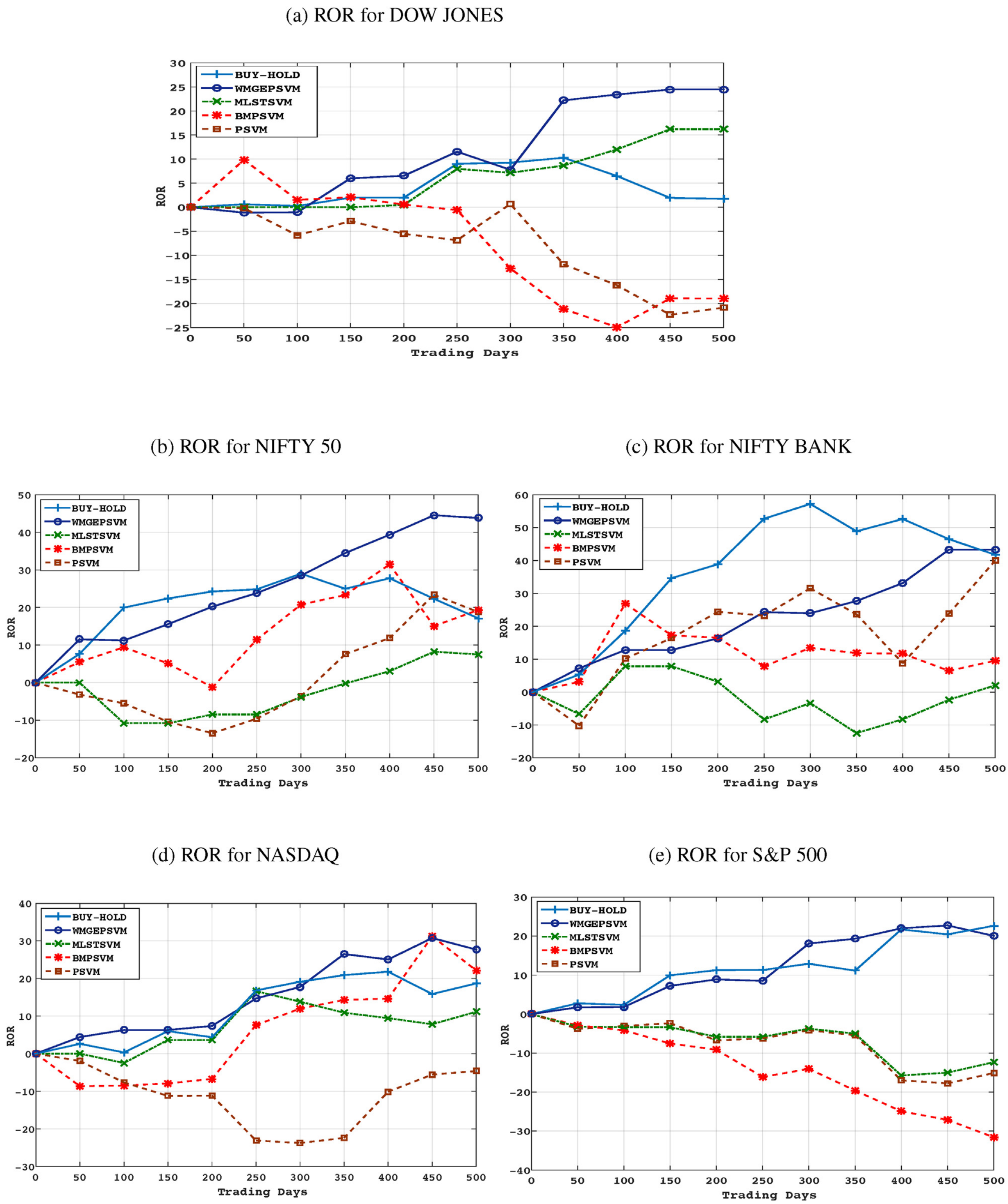


Fig. 3. ROR on 500 trading days testing data of different stock indices obtained by different trading systems without feature selection.

3.4. Training and parameter selection

The proposed trading system is trained by diving the whole data set into ten overlapping training-testing sets as depicted in Fig. 2. In this work, we use the walk forward (sliding window) testing approach which is extensively used in evaluating the performance of time-series forecasting system [32,31]. For each data set, 1500 consecutive days of data is considered as the *training set* and is used for training the classifier and feature pruning task, and subsequent 50 days of trading data is used for testing the performance of the proposed hybrid model. For the training and testing purpose in the net window, the training data is shifted 50 days forward by adding next 50 days and removing first 50 day so that the training sample size remain 1500 only. The testing is done on next 50 days. This process is repeated 10 times. In this way, data of 500 consecutive trading days is used as *testing data* for testing the performance of

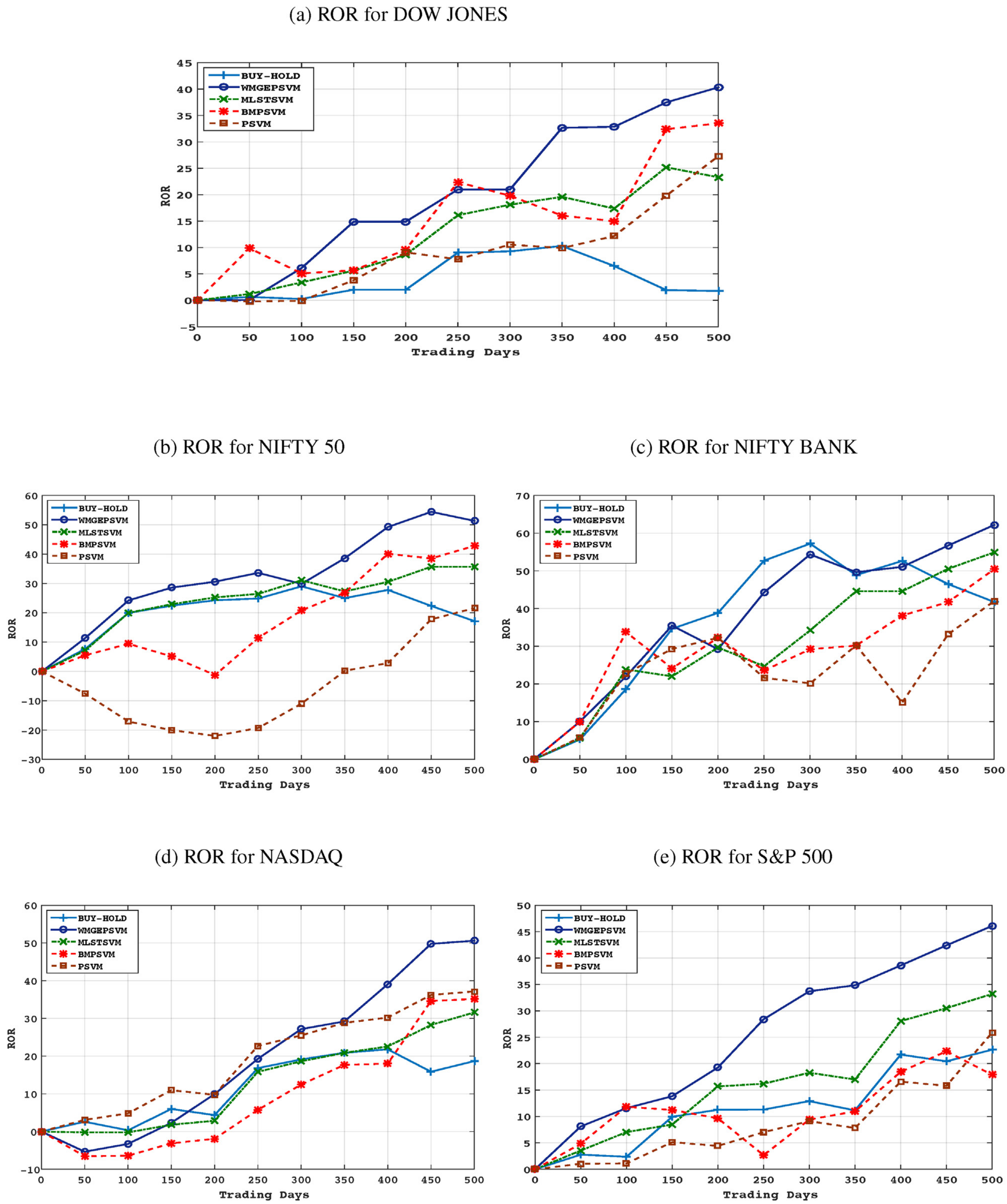


Fig. 4. ROR on 500 trading days testing data of different stock indices obtained by different hybrid trading systems.

the proposed financial trading system. The training set is utilized as input to the RF algorithm to rank all the input variables according to their importance. Then, these variables are added iteratively to WGEPSVM according to their ranks, i.e., a feature with the highest rank is added first and then the second highest rank feature is added, and so on. In the next step, WGEPSVM is trained with the training set. During the training phase, the optimization of regularization parameter ν of WGEPSVM is critical. Hence, optimal value of ν is determined based on trade-offs between generalization ability and forecasting accuracy. Regularization parameter ν is optimized by 5-fold Cross Validation (5-CV) method [24], [14], [43]. In 5-fold CV validation, training set is divided into five subsets (of equal size). One set is considered to be validation data and the remaining four subsets are considered for training. The value of the parameter ν

Table 2
Experimental results on testing data of different stock indices obtained by BUY-HOLD, PSVM, BMPSVM, LSTSVM and WMGEPSVM (best results are shown in bold).

Index name		BUY-HOLD	PSVM	BMPSVM	LSTSVM	WMGEPSVM
DOW JONES	ROR	1.75	−20.84	−18.96	16.20	**24.46**
	MDD		−37.68	−5.8	**−0.81**	−6.91
	PP		0.46	0.34	**0.83**	0.52
NIFTY BANK	ROR	41.75	39.99	9.55	2.03	**43.30**
	MDD		−53.55	−16.05	−12.39	**−7.83**
	PP		0.51	0.37	0.44	**0.61**
NIFTY 50	ROR	17.13	18.79	19.34	7.51	**43.81**
	MDD		−34.35	−14.52	−11.16	**−1.95**
	PP		0.46	0.42	**0.66**	0.60
NASDAQ	ROR	18.69	−4.63	22.12	11.13	**27.67**
	MDD		−49.36	−14.52	−11.25	**−5.51**
	PP		0.45	0.42	0.50	**0.52**
S&P 500	ROR	**22.65**	−15.09	−31.57	−12.32	18.30
	MDD		−25.10	−7.39	−10.71	**−3.41**
	PP		0.48	0.33	0.31	**0.57**

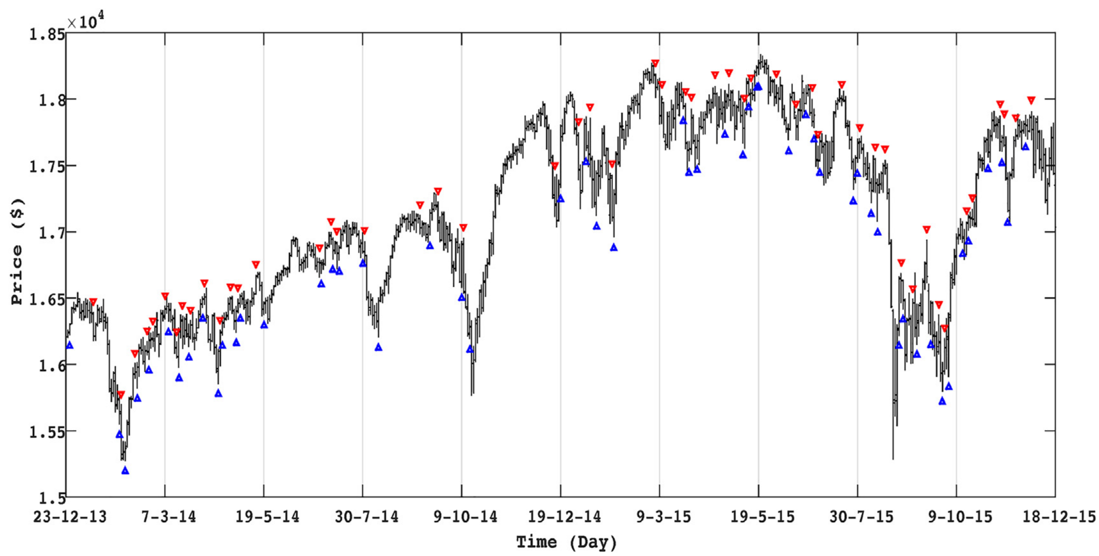


Fig. 5. Trading log ("▲": BUY, "▼": SELL) of DOW JONES for 500 trading days from December 23, 2013 to December 18, 2015 using RF-WMGEPSVM.

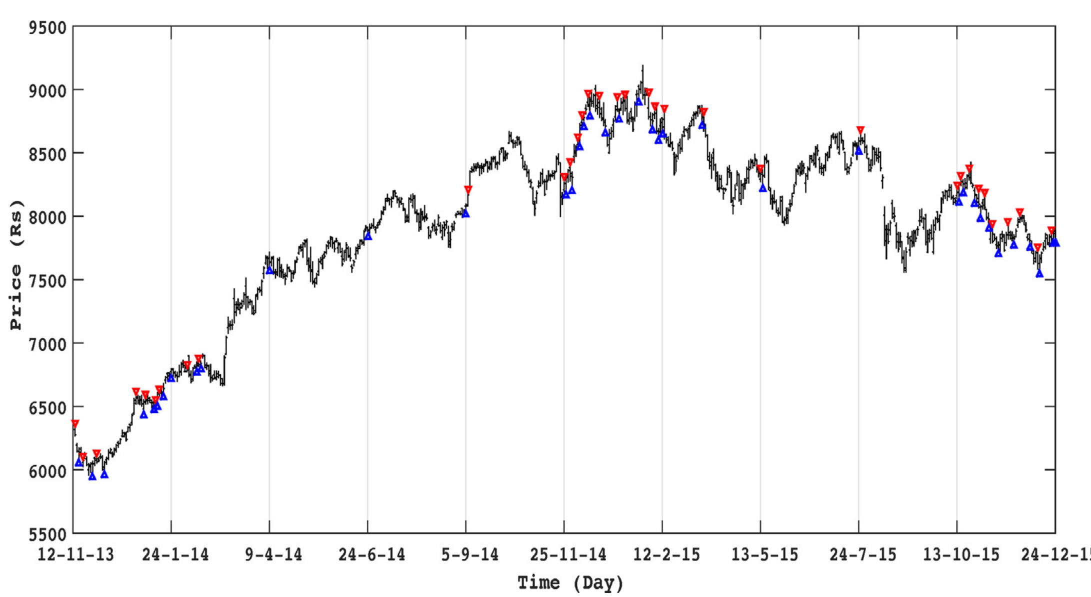


Fig. 6. Trading log ("▲": BUY, "▼": SELL) of NIFTY 50 for 500 trading days from November 12, 2013 to December 24, 2015 using RF-WMGEPSVM.

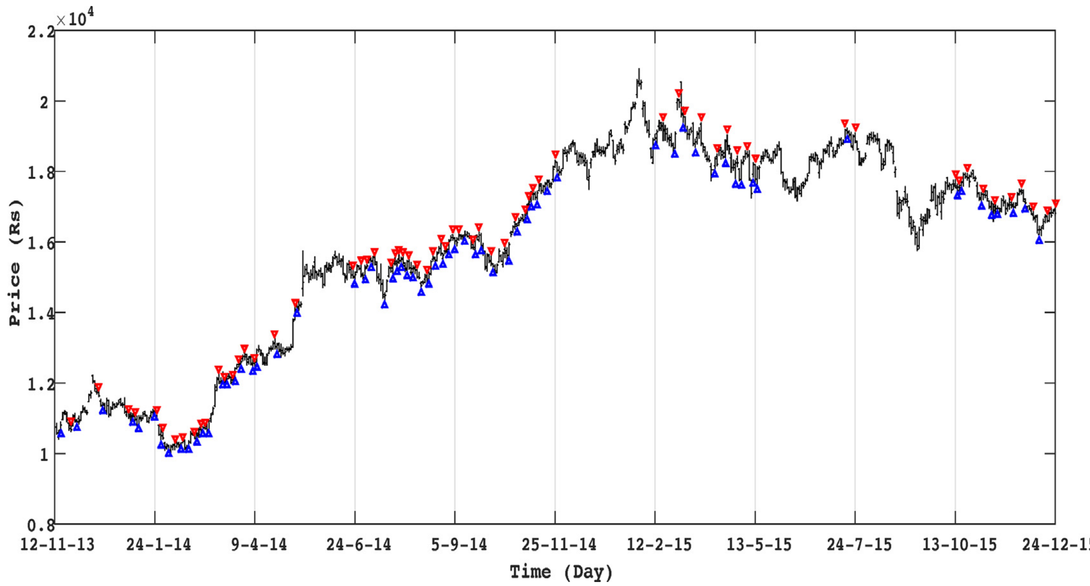


Fig. 7. Trading log ("▲": BUY, "▼": SELL) of NIFTY BANK for 500 trading days from November 12, 2013 to December 24, 2015 using RF-WMGEPSVM.

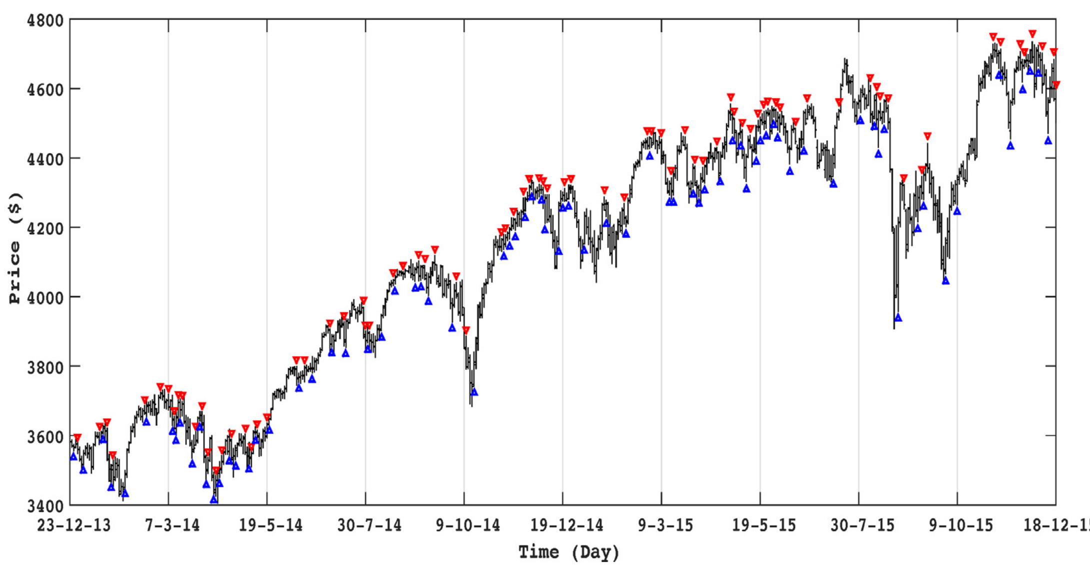


Fig. 8. Trading log ("▲": BUY, "▼": SELL) of NASDAQ for 500 trading days from December 23, 2013 to December 18, 2015 using RF-WMGEPSVM.

with the maximum average validation accuracy is used as the final recommended value of ν.

4. Experiment and discussions

The pertinence and effectiveness of proposed RF-WGEPSVM trading system have been tested on five index future contract data, viz., Dow Jones, NASDAQ, S&P 500, Nifty 50 and Nifty Bank Index. This data set has been carefully chosen as it is the representative of major financial markets around the globe. The time period for the empirical study is also chosen in such a manner that all the possible market scenarios are observed in the selected duration. The data sets used to conduct the experiment are collected from Quandl (https://www.quandl.com) for the period from January 2007 to December 2015. In this experiment, the total transaction cost is set to 0.05% of the transaction. For expediency to compare with other approaches all open positions are squared off at the end of each trading period of 50 days. All numerical experiments were performed on a computer with 2.5 GHz with 4 GB RAM and implemented in MATLAB. For the comparison under the homogeneous environment, same experimental setup was used to implement PSVM, RF-PSVM, BMPSVM, RF-BMPSVM, MLSTSVM and RF-MLSTSVM techniques. The proposed trading system is first tested without applying feature pruning. The proposed trading strategies are assessed on the basis of performance criteria (discussed in Section 2.1). Fig. 3 shows ROR of different all five index future for 500 consecutive trading days (10 out of sample data sets of 50 trading days each). Each point on this plot represents ROR for 50 trading days.

The ROR in fluctuant and range bound scenario for DOW JONES future for 500 trading days is depicted in Fig. 3a, WMGEPSVM achieved positive return in 7 out of 10 sample data sets with 24.46 % ROR while the market yields is only about 1.75% with negative ROR in five testing sets among the ten out of sample data sets. The PSVM and BMPSVM based trading strategies are found to loose 18.96% and 20.84%, respectively.

Fig. 3b and c shows the ROR in fluctuant and uptrend markets for NIFTY 50 and NIFTY BANK for 500 trading days, respectively. Both

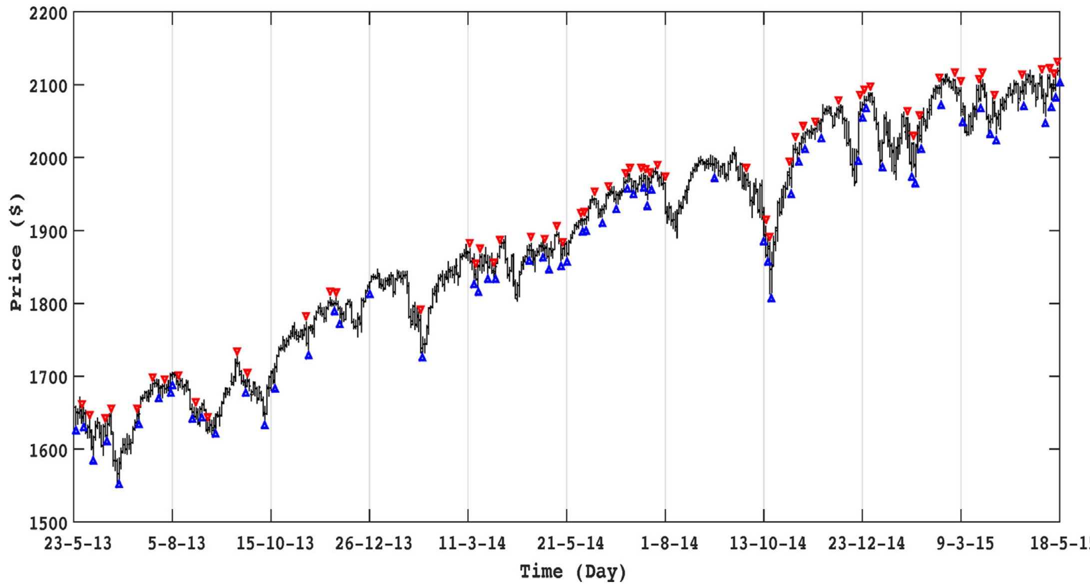


Fig. 9. Trading log ("▲": BUY, "▼": SELL) of S&P 500 for 500 trading days from May 23, 2013 to May 18, 2015 using RF-WMGEPSVM.

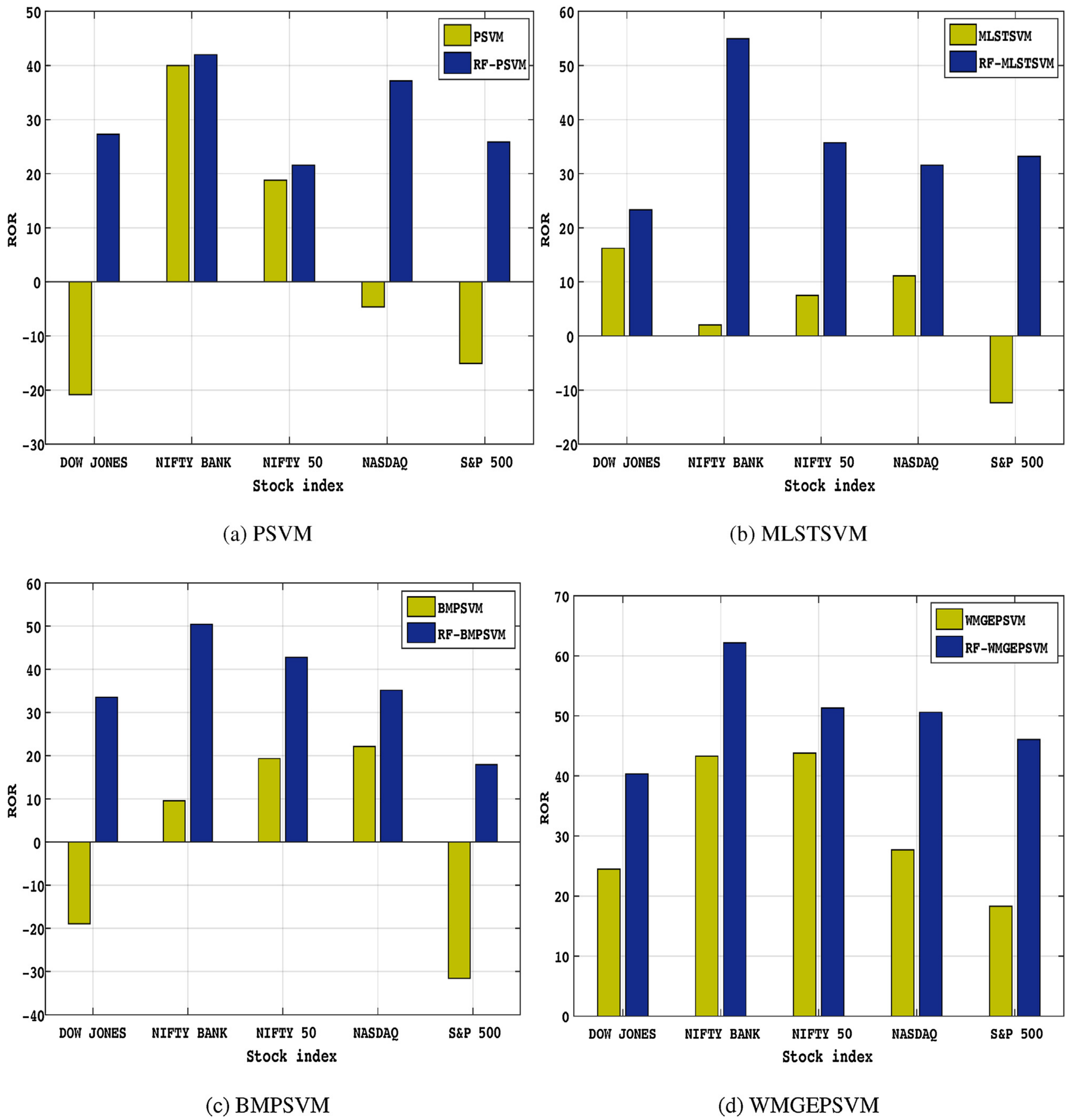
Table 3
Experimental results on testing data of different stock indices obtained by RF-PSVM, RF-BMPSVM, RF-LSTSVM and RF-WMGEPSVM (best results are shown in bold).

Index name		BUY-HOLD	RF-PSVM	RF-BMPSVM	RF-LSTSVM	RF-WMGEPSVM
DOW JONES	ROR	1.75	27.30	33.53	23.30	**41.72**
	MDD		−6.63	−6.52	−6.84	**−5.84**
	PP		0.46	0.37	**0.61**	0.60
NIFTY BANK	ROR	41.75	10.53	50.43	54.91	**63.13**
	MDD		−11.87	−13.86	−9.33	**−5.27**
	PP		0.44	0.43	0.54	**0.57**
NIFTY 50	ROR	17.13	21.57	42.79	35.70	**50.26**
	MDD		−10.69	−6.80	−5.82	**−2.84**
	PP		0.55	0.45	0.57	**0.58**
NASDAQ	ROR	18.69	37.15	35.14	31.57	**51.71**
	MDD		−6.51	−7.24	−5.09	**−4.63**
	PP		0.46	0.46	0.58	**0.60**
S&P 500	ROR	22.65	25.88	17.89	33.20	**40.44**
	MDD		−5.39	−29.38	−8.28	**−5.40**
	PP		**0.63**	0.42	0.60	0.60

indices show positive growth in the first 300 days, but the growth was negative between next 300 to 500 trading days. In such realistic market scenario, the proposed WMGEPSVM has shown 43.81% and 43.30% gain, respectively, while the market gains about 17.13% and 42.25% for NIFTY 50 and NIFTY BANK, respectively. It can be observed that WMGEPSVM is also able to make a profit while both indices lost in the 7th, 9th and 10th out of the sample test data set. PSVM, BMPSVM and LSTSVM trading strategies have also shown positive ROR for both the indices.

For NASDAQ and S&P 500 index futures, the ROR in fluctuant and bull markets are plotted in Fig. 3d and e. The corresponding market returns for these indices are about 17.69% and 22.65%, respectively. WMGEPSVM is having positive ROR in 9 out of 10 sample data sets with 27.67% returns for NASDAQ data set. For S&P 500 future, the proposed WMGEPSVM is also able to make positive ROR in 8 out of 10 sample data sets with 18.30% profit. For S&P 500, it is also noted that PSVM, BMPSVM, and LSTSVM based trading systems lost 15.09%, 31.57% and 12.32% respectively.

The other important measure used to evaluate the trading system is PP and MDD. A trading system with larger MDD would lead to investment deterioration since investors would lose money as well as confidence in the performance of the strategy. Therefore, a trading system with a smaller MDD is considered a more reliable. The detailed statistical results are reported in Table 2. WMGEPSVM performs better than BUY-HOLD, PSVM, BMPSVM and LSTSVM trading systems in terms of ROR, MDD and PP. The trading system using WMGEPSVM has achieved highest ROR for four out of five market indices. S&P 500 is the only indices for that the market return is higher than WMEPSVM. The trading system with WMGEPSVM has lower MDD for four indices than other trading systems. NIFTY BANK is the only index future for which LSTSVM has the lowest MDD. WMEPSVM is also achieved the highest PP for NIFTY BANK, NASDAQ and S&P 500. It is interesting to note that WMGEPSVM is the only trading system which is profitable for all five stock index futures. This is an indication of the fact that the WMGEPSVM is capable of capturing the uptrend, downside and range bound market trends scenarios with higher degree of success. After analyzing the significant better performance of the proposed WMGEPSM using all the indicators, the experiment is extended to make a hybrid trading system by amalgamating all four predictive algorithms with RF (named RF-PSVM, RF-BMPSVM, RF-MLSTSVM and RF-WMGEPSVM) algorithm. The results of hybrid Trading systems for all five stock indices on the same test data set (considered in the first case) are depicted in Fig. 4.



(a) PSVM (b) MLSTSVM

(c) BMPSVM (d) WMGEPSVM

Fig. 10. Demonstrating performance improvement by applying RF with different prediction models on all five stock indices.

It is observed from Fig. 4 that all hybrid systems achieve higher ROR on average as compared to BUY-HOLD strategy. Furthermore, all multi-class methods (RF-BMPSVM, RF-MLSTSVM, and RF-WMGEPSVM) produced BUY/SELL/HOLD trading signals that have shown better performance than RF-PSVM which is a “BUY-SELL” trading system.

The ROR acquired by all hybrid trading systems for DOW JONES index are plotted in Fig. 4a. RF-WMGEPSVM has produced positive ROR in all ten out of sample data sets while BUY-HOLD strategy lost in five out of sample sets (second, forth, eighth, ninth and tenth). It is also observed Fig. 4a that RF-WMGEPSVM has returned the highest profit (up to 41.72%) followed by RF-BMPSVM with ROR of 33.53% while BUY-HOLD strategy returns about 1.75%.

Fig. 4b and c shows the ROR attained by all hybrid decision support systems considered in this study for fluctuant and uptrend markets for NIFTY 50 and NIFTY BANK indices respectively. RF-WMGEPSVM shows the positive ROR in all ten out of sample sets however, the market lost in three out of sample (7th, 9th and 10th) test data set for both NIFTY 50 and NIFTY BANK indices.

The ROR for NASDAQ index is shown in Fig. 4d. The proposed hybrid trading system using RF-WMGEPSVM gains positive return in all ten out of sample data sets, while the market lost in five out of sample sets (2nd, 4th, 8th, 9th and 10th testing phase).

Fig. 4e shows the ROR obtained by different hybrid trading support systems for S&P 500 index future. As in the case of NASDAQ index, RF-WMEPSVM has the positive return in all ten out of sample sets, while the market lost in three out of sample sets (2nd, 7th and 9th test sample set).

For a visual illustration of the effectiveness of the RF-WMGEPSVM based trading system, trading logs using the RF-WMGEPSVM are plotted for all five indices. In the trading log, Date is plotted on the *X*-axis and the bar chart of the index future value is plotted on *Y*-axis. A blue "▲" indicates that the proposed decision support system generates a BUY signal, and a red "▼" indicates that a SELL signal is generated by the proposed decision support system.

From Fig. 4a, it is clearly observed that the proposed RF-WMGEPSVM based trading system outperforms other trading system for all five market indices. The significance of the RF-WMGEPSVM technique lies in the fact that it can capture the major change in the index movements, are shown in Figs. 5–9.

The detailed statistical results for all hybrid trading systems are summarized in Table 3. The performances of trading systems have significantly improved by applying the RF technique in conjunction with various predictive methods. The experimental results indicate that RF-WMGEPSVM based trading system outperforms all other trading systems in terms of ROR. The trading systems making use of RF-WMGEPSVM performs better as compared to BUY-HOLD, PSVM, BMPSVM, and MLSTSVM based trading systems on the basis of ROR, MDD, and PP. RF-WMGEPSVM has achieved the lowest MDD for 3 out of 5 index future data sets, and achieved the highest PP for 3 out of 5 index future data sets. The trading system with RF-MLSNHSVM is able to achieve the lowest MDD and the highest PP for 2 out of 5 index future data sets. All hybrid trading systems are found to outperform the BUY-HOLD strategy, which indicates the superiority of machine learning based trading systems over the naive BUY-HOLD investment strategy (refer Table 3).

Finally, all four trading systems and their hybrid versions are compared in Fig. 10, on the basis of their respective ROR for all five stock indices. From Fig. 10, it is noted that all the hybrid trading systems perform significantly better than their non-hybrid versions. Moreover, it also noted that all the hybrid trading system are able to achieve positive ROR in all five markets. This clearly indicates the reliability and superiority of hybrid trading system over the non-hybrid systems.

5. Conclusion

In this study, decision support systems for making the automated trading decision are proposed. For this purpose, WMGEPSVM algorithm is amalgamated with RF technique. The trading rules have been constructed using a "Buy/Hold/Sell" signal, which are received on daily close prices of the stock market. The proposed trading systems are tested over 500 days daily data of DOW JONES, NIFTY 50, NIFTY BANK, NASDAQ, and S&P 500 index futures.

The performance is evaluated on the basis of various performance metrics, viz., ROR, PP and MDD ratio. The experimental results indicate that WMGEPSVM performs significantly better than Buy-HOLD approach, PSVM, BMPVM and MLSTVM. This is due to the fact that the classification of unbalanced data classes can be effectively handled by the WMGEPSVM. The input variable is some of the widely used technical indicators and oscillators in the financial market by experts. For the purpose of selecting more informative indicators, RF technique is used. Use of RF technique has demonstrated notable improvements on the overall performance of all predictive methods (without using feature pruning).

After applying RF technique, it is observed that dimensionality reduction not only reduces the computational complexity of the models but also improves the performance of the financial trading systems. Empirical findings also suggest the superiority of the proposed "BUY/SELL/HOLD" signals based trading systems (RF-WMGEPSVM, RF-PSVM, RF-BMPSVM, and RF-LSTSVM) as compared to the BUY-HOLD investment strategy and RF-PSVM based trading system. The proposed RF-WMGEPSVM based trading system is found to outperform all other trading systems based on other classifiers considered in this study in terms of ROR, PP, and MDD in bull, bearish and sideways market scenario.

The results obtained in this study are promising and may be useful for portfolio making in the financial market.

References

[1] Y.S. Abu-Mostafa, A.F. Atiya, Introduction to financial forecasting, Appl. Intell. 6 (3) (1996) 205–213.
[2] G. Deboeck, Trading on the Edge: Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets, vol. 39, John Wiley & Sons, 1994.
[3] S.C. Blank, "Chaos" in futures markets? A nonlinear dynamical analysis, J. Futures Markets 11 (6) (1991) 711–728.
[4] C.-F. Tsai, Y.-C. Lin, D.C. Yen, Y.-M. Chen, Predicting stock returns by classifier ensembles, Appl. Soft Comput. 11 (2) (2011) 2452–2459.
[5] L. Cao, F. Tay, Support vector machine with adaptive parameters in financial time series forecasting, IEEE Trans. Neural Netw. 14 (6) (2003) 1506–1518.
[6] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (5) (1989) 359–366.
[7] S. Haykin, R. Lippmann, Neural networks, a comprehensive foundation, Int. J. Neural Syst. 5 (4) (1994) 363–364.
[8] L. Yu, S. Wang, K.K. Lai, A neural-network-based nonlinear metamodeling approach to financial time series forecasting, Appl. Soft Comput. 9 (2) (2009) 563–574.
[9] G. Grudnitski, L. Osburn, Forecasting S&P and gold futures prices: an application of neural networks, J. Futures Markets 13 (6) (1993) 631–643.
[10] W. Leigh, R. Purvis, J.M. Ragusa, Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support, Decis. Support Syst. 32 (4) (2002) 361–377.
[11] G.P. Zhang, An investigation of neural networks for linear time-series forecasting, Comput. Oper. Res. 28 (12) (2001) 1183–1202.
[12] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[13] V.N. Vapnik, V. Vapnik, Statistical Learning Theory, vol. 1, Wiley, New York, 1998.
[14] L. Yu, H. Chen, S. Wang, K.K. Lai, Evolving least squares support vector machines for stock market trend mining, IEEE Trans. Evol. Comput. 13 (1) (2009) 87–102.
[15] W. Huang, Y. Nakamori, S.-Y. Wang, Forecasting stock market movement direction with support vector machine, Comput. Oper. Res. 32 (10) (2005) 2513–2522.
[16] K.-j. Kim, Financial time series forecasting using support vector machines, Neurocomputing 55 (1) (2003) 307–319.
[17] D. Kumar, S.S. Meghwani, M. Thakur, Proximal support vector machine based hybrid prediction models for trend forecasting in financial markets, J. Comput. Sci. 17 (2016) 1–13.
[18] J. Patel, S. Shah, P. Thakkar, K. Kotecha, Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques, Expert Syst. Appl. 42 (1) (2015) 259–268.
[19] C.-Y. Yeh, C.-W. Huang, S.-J. Lee, A multiple-kernel support vector regression approach for stock market price forecasting, Expert Syst. Appl. 38 (3) (2011) 2177–2186.
[20] Y. Pan, Z. Xiao, X. Wang, D. Yang, A multiple support vector machine approach to stock index forecasting with mixed frequency sampling, Knowl. Based Syst. 122 (2017) 90–102.
[21] C.-F. Huang, A hybrid stock selection model using genetic algorithms and support vector regression, Appl. Soft Comput. 12 (2) (2012) 807–818.
[22] N. Verma, S. Das, N. Srivastava, Multiple kernel support vector regression for pricing nifty option, Int. J. Appl. Math. Res. 4 (4) (2015) 488–495.
[23] P.-C. Chang, J.-L. Wu, J.-J. Lin, A Takagi–Sugeno fuzzy model combined with a support vector regression for stock trading forecasting, Appl. Soft Comput. 38 (2016) 831–842.
[24] L. Cao, K. Chua, W. Chong, H. Lee, Q. Gu, A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine, Neurocomputing 55 (1) (2003) 321–336.
[25] L.-J. Kao, C.-C. Chiu, C.-J. Lu, J.-L. Yang, Integration of nonlinear independent component analysis and support vector regression for stock price forecasting, Neurocomputing 99 (2013) 534–542.
[26] M.-C. Lee, Using support vector machine with a hybrid feature selection method to the stock trend prediction, Expert Syst. Appl. 36 (8) (2009) 10896–10904.
[27] R. Tsaih, Y. Hsu, C.C. Lai, Forecasting S&P 500 stock index futures with a hybrid AI system, Decis. Support Syst. 23 (2) (1998) 161–174.
[28] N. Baba, N. Inoue, Y. Yanjun, Utilization of soft computing techniques for constructing reliable decision support systems for dealing stocks, in:

Proceedings of the 2002 International Joint Conference on Neural Networks, 2002, IJCNN'02, vol. 3, IEEE, 2002, pp. 2150–2155.
[29] S.S. Alexander, Price movements in speculative markets: trends or random walks, Ind. Manag. Rev. (pre-1986) 2 (2) (1961) 7.
[30] R. Hudson, M. Dempsey, K. Keasey, A note on the weak form efficiency of capital markets: the application of simple technical trading rules to UK stock prices – 1935 to 1994, J. Bank. Finance 20 (6) (1996) 1121–1132.
[31] Y.-K. Kwon, B.-R. Moon, A hybrid neurogenetic approach for stock forecasting, IEEE Trans. Neural Netw. 18 (3) (2007) 851–864.
[32] Q. Wen, Z. Yang, Y. Song, P. Jia, Automatic stock decision support system based on box theory and SVM algorithm, Expert Syst. Appl. 37 (2) (2010) 1015–1022.
[33] A. Smola, V. Vapnik, Support vector regression machines Advances in Neural Information Processing Systems, vol. 9, 1997, pp. 155–161.
[34] D. Kumar, M. Thakur, Weighted multicategory nonparallel planes SVM classifiers, Neurocomputing 211 (2016) 106–116.
[35] G.M. Fung, O.L. Mangasarian, Multicategory proximal support vector machine classifiers, Mach. Learn. 59 (1–2) (2005) 77–97.
[36] D. Tomar, S. Agarwal, A comparison on multi-class classification methods based on least squares twin support vector machine, Knowl. Based Syst. 81 (2015) 131–147.
[37] J.D. Schwager, A Complete Guide to the Futures Markets: Fundamental Analysis, Technical Analysis, Trading, Spreads, and Options, John Wiley & Sons, 1984.
[38] R. Rhea, The Dow Theory: An Explanation of Its Development and an Attempt to Define Its Usefulness as an Aid in Speculation, Fraser Publishing Company, 1993.
[39] J.J. Murphy, Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications, Penguin, 1999.
[40] M.J. Pring, Technical Analysis Explained: The Successful Investor's Guide to Spotting Investment Trends and Turning Points, McGraw-Hill Professional, 2002.
[41] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[42] C.-L. Huang, C.-Y. Tsai, A hybrid SOFM–SVR with a filter-based feature selection for stock market forecasting, Expert Syst. Appl. 36 (2 Pt 1) (2009) 1529–1539.
[43] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, R. Tibshirani, The Elements of Statistical Learning, vol. 2, Springer, 2009.