International Conference on Automation, Control Engineering and Computer Science (ACECS)
Proceedings of Engineering and Technology – PET
Vol.21pp.91-95

Stock price prediction based on SVM : The impact of the stock market indices on the model performance

Anass NAHIL
Laboratory of Innovative Technologies (LTI)
ENSA of Tangier
Abdelmalek Essaadi University
anassnahil@gmail.com

Abdelouahid Lyhyaoui
Laboratory of Innovative Technologies (LTI)
ENSA of Tangier
Abdelmalek Essaadi University
lyhyaoui@gmail.com

***Abstract*— The challenge of stock forecasting is appealing because a small forecasting improvement can increase profit significantly. However, the volatile nature of the stock market makes it difficult to apply linear models, simple time-series or regression techniques. Consequently, support vector machine (SVM) has become a good alternative. It is a popular tool in time series forecasting for the capital investment industry. This machine learning technique which is based on a discriminative classifier algorithm, forecasts more accurately the financial data. By examining the stock price of 5 Moroccan banks, the experiment shows that the SVM can perform better when we add the global evolution of the market to the independent variables. To express the global evolution of the market, three indices of the Casablanca Stock Exchange are used : MASI, MADEX and Banks Sector Index.**

Keywords: Stock price prediction; Financial time series; Support vector machines; the Moroccan Stock Market; Casablanca Stock Exchange.

I. INTRODUCTION

Stock price forecasting is valuable for investors. It tells out the investment opportunities. Unlike other methods which are concerned with company fundamental analysis, in our approach, the independent variables are derived from the stock itself.

Research efforts have been made to find superior forecasting methods, and to enhance existing ones. Many studies have found that univariate time series models are an accurate forecasting models [1], [2], [3]. However, their accuracy require a linear and not very volatile data. The financial data doesn't satisfy those conditions. Indeed, the stock prices are random and cannot be linearly predicted. Sharda et al.[4], Haykin [5] and Zhang et al. [6] studies' have shown that, compared to traditional statistical models, neural networks describes more accurately the movement of financial time series. Those Machine learning techniques have been successfully used for modeling financial time series. Hall JW. [7] have used neural networks in an adaptive selection of U.S stocks. S.Kim et al. [8] have applied the probabilistic neural networks to a stock market index. Saad et al. [9] have predicted the trend of a stock market using time delay, recurrent and probabilistic neural networks.

The support vector machine [10], used in this paper, is a neural network technique that has been widely used in stock price predictions. It is a popular tool in time series forecasting [11], [12], due to its good generalization performance, the absence of local minima and the sparse representation of solution. The training of SVM is equivalent to solving a linearly constrained convex quadratic programming problem and the solution is always absent from local minima, it is global and optimal. In the beginning, SVMs have been developed for pattern recognition [13]. However, After the introduction of the Vapniks ϵ-insensitive loss function, SVMs have been extended to solve non-linear regression estimation problems [14], [15]. In this paper, we apply SVM regression method, to construct the prediction models for forecasting the five major Moroccan banks stock price.

Many studies have been completed in the field of stock price prediction using SVM and machine learning techniques. Their purpose was to predict the future behavior of the stock price in different stock exchange markets. The ANN model have been used by Refenes et al. [16], Tsibouris et al. [17] and Steiner et al. [18]. Wittkemper et al. [19] and Shazly et al. [20] have used neural network along with genetic algorithmic (hybrid system). Tay and Cao [21] have formulated a pricing model for futures in US market using SVMs. Gestel et al. [22] have used LS-SVM for T-Bill rate and stock index pricing in US and German market. Chen et al. [23] have applied SVM and back propagation neural networks. Tasi [24] has studied bankruptcy using SVM and compare it to neural network. Tay et al. [25] have studied the feasibility of applying SVM for financial time series forecasting using the Chicago mercantile markets data sets. Ince et al. [26] have applied a short term forcasting with SVM to the stock price prediction. Rudra et al. [27] have used A Nave SVM-KNN based stock market trend reversal analysis for Indian benchmark indices.

The purpose of this paper is to enhance the performance of SVM by the introduction of the global evolution of the market. This global evolution is reflected by three major Moroccan stock indices which are : MASI, MADEX and Banks Sector Index.

This paper consists of five sections. Section 2 presents the principles of SVMs regression. Section 3 exhibits the procedures involving data set selection, data preprocessing and kernel function selection. The experimental results followed by the conclusions drawn from this study are presented in section 4.

ISSN 2356-5608

II. SVM REGRESSION THEORY

Given a set of data points $\{(x_1, y_1), ..., (x_l, y_l)\}$, $x_i \in X \subset R^n$ is the input vector, l is the total number of data patterns and $y_i \in R$ is the i^{th} value of the dependent variable. The estimating function f is approximated using the following :

$$y = f(x) = (w.\phi(x)) + b \tag{1}$$

where $\phi(x)$ is the high dimensional feature space which is non-linearly mapped from the input space x. The coefficients w and b are estimated by minimizing the risk function :

$$R_{SVM}(C) = C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|w\|^2 \tag{2}$$

$$L_\varepsilon(d_i, y_i) = \begin{cases} |d-y| - \epsilon & |d-y| \geq \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

L_ϵ is the extension of ϵ-insensitive loss function.
Considering the slack variables (ζ_i, ζ_i^*), the problem can be reformulated as :

$$\text{(P1) : Minimize C } \textstyle\sum_{i=1}^{l}(\zeta_i + \zeta_i^*)\frac{1}{2}\|w\|^2$$

$$\text{Subject to} \begin{cases} y_i - w.\phi(x_i) - b \leq \epsilon + \zeta_i \\ y_i w.\phi(x_i) - b \leq \epsilon + \zeta_i^* \\ \zeta_i \geq O \\ \zeta_i^* \geq O \end{cases} \quad i = 1, 2..., l \tag{4}$$

C is an user specified constant. The solution of (P) using primal dual method leads to the following new problem : Determine the Lagrange multipliers $\{\alpha_i\}_{i=1}^{l}$ and $\{\alpha_i^*\}_{i=1}^{l}$ solving the following problem :

$$\begin{gathered}\text{(P2) : Maximize } Q(\alpha_i, \alpha_i^*) = \textstyle\sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \\ -\epsilon \textstyle\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) \\ -\frac{1}{2}\textstyle\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) \\ \text{Subject to} \begin{cases} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \\ 0 \leq \alpha_i \leq C \\ 0 \leq \alpha_i^* \leq C \end{cases} \quad i = 1, 2..., l \end{gathered} \tag{5}$$

C is a constant and K is the Mercer Kernel :

$$\begin{gathered} K : X \times X \to R \\ (x, z) \to \phi(x).\phi(z) \end{gathered} \tag{6}$$

The typical examples of Kernel function are as follows:
Linear : $K(x_i, x_j) = x_i^T x_j$;
Polynomial : $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$;
Radial basis function (RBF) :
$K(x_i, x_j) = exp(-\gamma\|x_i - x_j\|^2), \gamma > 0$;
Sigmoid : $K(x_i, x_j) = tanh = (\gamma x_i^T x_j + r)$.
The kernel parameters (γ, r and k) should be carefully chosen. Those parameters control the complexity of the final solution by defining the structure of the high dimensional feature space $\phi(x)$. The solution of the primal (P2) yields : $w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\phi(x_i)$ and b is calculated using Karush-Kuhn-Tucker conditions (KKT) :

$$\begin{cases} \alpha_i(\epsilon + \zeta_i - y_i + w.\phi(x_i) + b) = 0 \\ \alpha_i^*(\epsilon + \zeta_i^* - y_i + w.\phi(x_i) + b) = 0 \\ (C - \alpha_i)\zeta_i = 0 \\ (C - \alpha_i^*)\zeta_i^* = 0 \end{cases} \tag{7}$$

Since $\alpha_i, \alpha_i^*, \zeta_i^* = 0$ for $\alpha_i^* \in (0, C)$, b can be computed as follows :

$$\begin{cases} b = y_i - w.\phi(x_i) - \epsilon & 0 \leq \alpha_i \leq C \\ b = y_i - w.\phi(x_i) - \epsilon & 0 \leq \alpha_i \leq C \end{cases} \tag{8}$$

For those α_i, α_i^* for which x_i corresponds to $0 \leq \alpha_i^* \leq C$ and $0 \leq \alpha_i \leq C$ are called support vectors. The number of support vectors is a function of ϵ, the larger the ϵ, the fewer the number of support vectors and thus the sparser the representation of the solution. However, a larger ϵ depreciate the model accuracy. In this regard, ϵ should be a compromise between the sparseness of representation and the closeness to data.
Considering the previous results, $f(x)$ can be computed as :

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b \tag{9}$$

III. PROPOSED METHODOLOGY

A. Experimental design

For each model in this paper, the SVMs regression forecasting follows the procedure in figure 1.

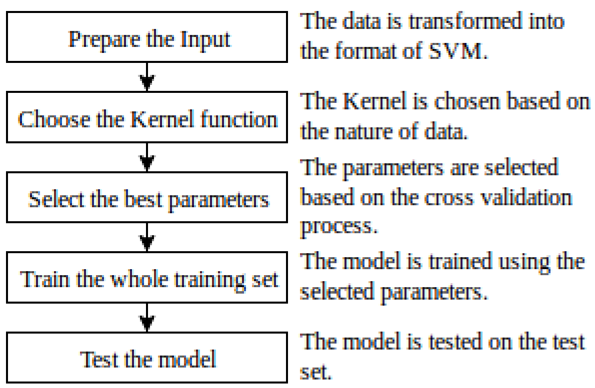

Fig. 1: The experimental design of the SVM regression

B. Data sets

Five stock prices of the five major Moroccan banks, collated from The Casablanca Stock Exchange, are examined in the experiment. They are : Banque Populaire (BP), Attijari Wafa Bank (AWB), Banque Marocaine pour le Commerce et l'Industrie (BMCI), BMCE Bank (BMCE) and CIH Bank (CIH). We have taken 1119 samples for each of the stocks mentioned above. The corresponding time period is from 2nd January, 2012 to 30th June, 2016. We use two-thirds of the research data points (the first 746 closing price) as the training data. The 373 remaining data points are used as the test data.

ISSN 2356-5608

	BCP	BMCI	AWB	BMCE	CIH
Min	178.75	506.00	300.00	137.00	185.00
Max	239.85	910.00	382.00	240.00	369.95
Mean	204.43	735.16	333.65	202.86	272.79
Median	198.00	765.00	332.08	206.10	271.15
SD	14.42	95.14	19.62	18.12	41.14

TABLE 1: Description of the data sets

The collected data consists of daily closing price. They are used as the data sets. Table 1 shows high price, low price, mean, median and standard deviation (SD) of the five stocks collected for our experiment. The calculated standard deviation shows that the stocks of the banks sector are relatively not very volatile which explains the good performances of the models used in this paper.

C. Data preprocessing

The original closing price is transformed into a five-day relative difference in percentage of price (RDP), as suggested by Thomason [28], in order to enhance the forecasting ability of the model. The transformed data are more symmetrical and follow more closely a normal distribution. One more transformed closing price is obtained by subtracting a 15-day exponential moving average (EMA15) from the closing price. The subtraction is performed to eliminate the trend. EMA15 is used to maintain the information contained in the original closing price since the application of the transformation of the closing price may remove some useful information. The output variable $RDP_{+5}^{EMA_3}$ is obtained after a smoothing of the closing price with a three-day exponential moving average. The smoothing enhances the prediction performance of SVM. The input and output variables are presented in table 2.

	Indicator	Calculation
Input	$\tilde{P}(i)$	$P(i) = EMA_{15}(i)$
	RDP_j	$\frac{P(i)-P(i-j)}{P(i-j)} * 100$
Output	$RDP_{+5}^{EMA_3}$	$\frac{EMA_3(i+5)-EMA_3(i)}{EMA_3(i)} * 100$

TABLE 2: Performance indicators

Giving P(i) the closing price of the day i, the n-day exponential moving average of the i$^{\text{th}}$ $EMA_N(i)$, mentioned in table 2, can be computed as follows :
$EMA_N(1)$ is the simple moving average calculated from the N previous observations, for $i > 1$:

$$\begin{aligned} EMA_N(i) = &\frac{2}{N+1}(P(i) - EMA_N(i-1)) \\ &+ EMA_N(i-1) \end{aligned} \tag{10}$$

The exponential moving averages reduce the lag by applying more weight to recent prices. This weight depends on the number of periods (N) in the moving average. Based on those performance indicators, two models are prepared. The first model (SVM model) applies the performance indicators (table 2) to the stock price of each bank. The difference in percentage of price is calculated for j=5,10,15 and 20 as recommended by Thomason [28]. The second model (SVM+ model) applies the same indicators to each bank then adds the three indices of the Casablanca stock exchange (MASI, MADEX and Banks Sector Index) to the input (independent variables). Since outliers may make it difficult to arrive to an effective solution, values beyond a set limit of standard deviation are selected as outliers. They are replaced with the closest marginal values. A limit of ± 2 SD is set for the RDP values and a limit ± 50 SD is set for indexes values. Another pre-processing technique used in this study is data scaling. In support vector machines, feature scaling improves the convergence speed of the algorithm [29]. For that reason, All the data points are scaled into the range of [0.9; 0.9].

D. Kernel function and parameters selection

A kernel is a function that satisfied the Mercers condition. The two typical kernels used in the Vapniks SVM for regression are the polynomial Kernel and the Gaussien Kernel. The Gaussien Kernel function performs well under general smoothness assumptions. Polynomial Kernel takes a longer time in training SVMs and gives inferior result compared to Gaussian Kernel. When training a model, the parameters that gives the best performance should be selected. Generally, in the SVR, those values of ϵ that produces the best result on the validation set of our data are selected. In the case of a Gaussien Kernel, we introduce two additional parameters: C and γ. A model selection must be done to identify good (C,γ) so that the classifier can accurately perform in the test data. We use a grid-search on C and γ based on a cross-validation process. The values of the parameters used in the selection process are illustrated in table 3.

C	γ
10	0.1
100	0.01
1000	0.001
10000	0.0001

TABLE 3: Grid-search parameters

E. Performance Criteria

The prediction accuracy is evaluated using the statistical metrics in table 4, namely, the normalized mean squared error (NMSE), mean absolute error (MAE), directional symmetry (DS) and the coefficient of determination R^2.

Metrics	Calculation
NMSE	$NMSE = \frac{1}{n\delta^2}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$ $\delta^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_i)^2$
MAE	$\frac{1}{n}\sum i = 1n\,\lvert x_i - \hat{x}_i\rvert^2$
DS	$DS = \frac{100}{n}\sum_{i=1}^{n}\delta_i$ $\delta_i = \begin{cases} 1 & (x_i - x_{i-1})(\hat{x}_i - \hat{x}_{i-1}) \geq 0 \\ 0 & otherwise \end{cases}$
R^2	$1 - \sum_{i=1}^{n}(x_i - \hat{x_i})^2 / \sum_{i=1}^{n}(x_i - \bar{x_i})^2$

TABLE 4: Calculation of the performance metrics

ISSN 2356-5608

IV. RESULTS AND DISCUSSION

The SVM model applied to the data sets let as draw the following figures. The selected model from the training set is applied to the test set. The figures illustrate the real values and the predicted values for the test set of each bank using SVM and SVM+. Globally, SVM+ performs better, but, in some data points, predicted values by SVM are closer to the real values. However, The prediction made by the SVM+ model remains more risk-averse. Indeed, the predicted values by SVM+ are, globally, bellow those predicted by SVM. This observation is due to the addition of the market global evolution to the model. This variable reflects the average evolution of the market which is more risk-averse.

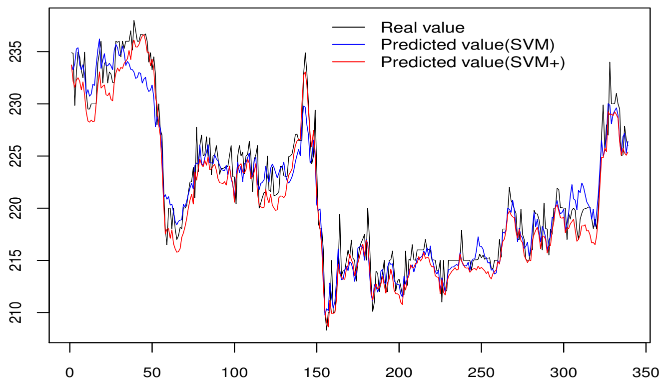

Fig. 2: Results of the fitting for BCP

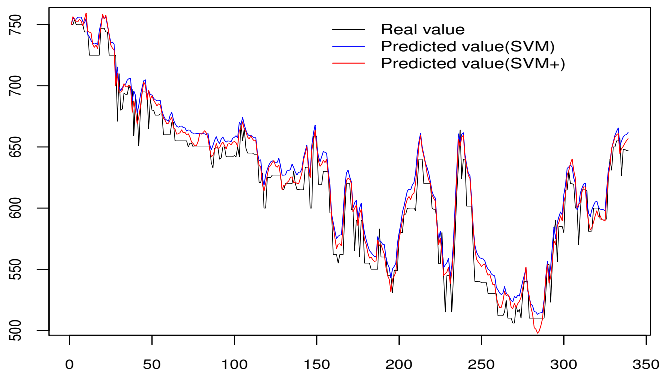

Fig. 3: Results of the fitting for BMCI

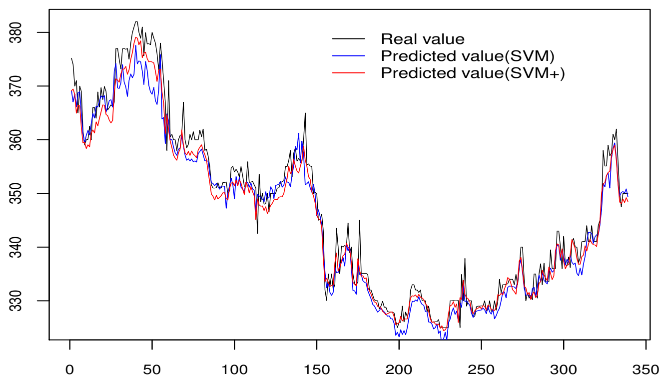

Fig. 4: Results of the fitting for AWB

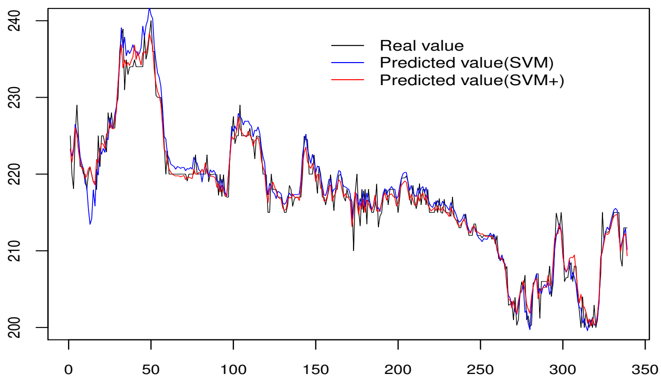

Fig. 5: Results of the fitting for BMCE

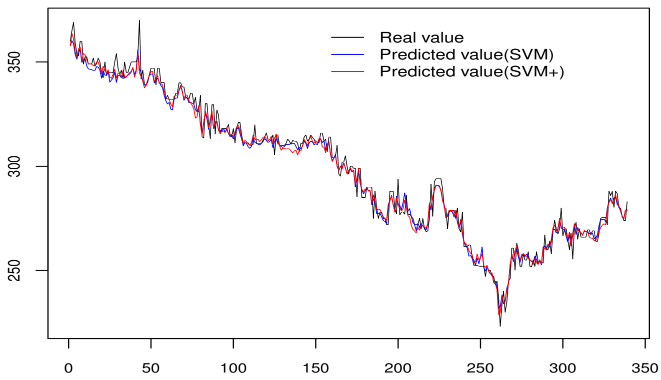

Fig. 6: Results of the fitting for CIH

The optimal values of the parameters of SVM chosen based on the validation set are given in Table 5. The results obtained are given in Table 6. Obviously, SVM+ converges to a better performance indicators on the test set. SVM+ provides a smaller NMSE and MAE and larger DS and R^2 than SVM+ in all the cases (tables 6 and 7).

	SVM		SVM+	
	C	γ	C	γ
BCP	0.00001	100	0.00001	1000
BMCI	0.00001	10	0.00001	10
AWB	0.001	1000	0.001	10
BMCE	0.0001	100	0.001	100
CIH	0.01	10	0.001	100

TABLE 5: The selected simulation parameters

	NMSE		MAE	
	SVM	SVM+	SVM	SVM+
BCP	0.002811	0.002585	1.208029	1.201884
BMCI	0.001260	0.001250	6.014052	5.976358
AWB	0.001837	0.001341	2.024262	0.269665
BMCE	0.002073	0.001759	1.185295	1.129769
CIH	0.001268	0.001016	3.297368	3.217085

TABLE 6: The converged indicators : NMSE and MAE

ISSN 2356-5608

	DS		R^2	
	SVM	SVM+	SVM	SVM+
BCP	65.19	65.78	95.24%	95.25%
BMCI	84.37	85.55	97.85%	97.86%
AWB	62.54	62.54	96.29%	96.79%
BMCE	69.62	70.50	95.92%	96.35%
CIH	66.37	67.55	98.13%	98.20%

TABLE 7: The converged indicators : DS and R^2

Reviewing results in regards to the NMSE, the model accuracy improvement[1] is the greatest and increases by 26.97% for AWB. For BMCI, the NMSE remains almost constant (table 8). This result is supported by the market capitalization of each bank that reflect the importance and economic weight of the bank in the market. A stock price that has an important market capitalization is more correlated to the global trend of the market, thus the information provided by the global evolution of the market contributes more significantly to the model improvement.

Accuracy improvement (AI) based on NMSE			
	SVM	SVM+	AI (in %)
BCP	0.002811	0.002585	8.04%
BMCI	0.001260	0.001250	0.82%
AWB	0.001837	0.001341	26.97%
BMCE	0.002073	0.001759	15.14%
CIH	0.001268	0.001016	19.90%

TABLE 8: The models accuracy improvement

V. CONCLUSIONS

The information contained in the trend of the stock market helps to improve the performance of the SVM regression. In order to evaluate the accuracy improvement in the model, this paper proposes to add the trend of the stock market to the input variables of the SVM model. The global evolution of the market is represented by the indices of the stock market and the sectoral index. The enhanced model provides a better fitting for the stock prices, particularly for those with an important market capitalization.

REFERENCES

[1] D.A. Bessler and J.A. Brandt. Forecasting livestock prices with individual and composite methods. Applied Economics, 13, 513-522, 1981.

[2] K.S. harris and R.M. Leuthold. A comparison of alternative forecasting techniques for live stock prices: A case study. North Central J. Agricultural Economics, 7, 40-50, 1985.

[3] J.H. Dorfman and C.S. Mcintosh. Results of a price forecasting competition. American J. Agricultural Economics, 72, 804-808, 1990.

[4] Sharda R, Patil RB. A connectionist approach to time series prediction: an empirical test. In: Trippi, RR, Turban, E, (Eds.), Neural Networks in Finance and Investing, Chicago: Probus Publishing Co., 1994.

[5] Haykin S. Neural networks: a comprehensive foundation. Englewood CliKs, NJ: Prentice Hall, 1999.

[6] Zhang GQ, Michael YH. Neural network forecasting of the British Pound=US Dollar exchange rate. Omega 1998, 26(4), 495-506.

[7] Hall JW. Adaptive selection of U.S. stocks with neural nets. In: GJ Deboeck (Ed.),Trading on the edge: neural, genetic, and fuzzy systems for chaotic financial markets. New York:Wiley, 1994.

[1] 1-the ratio of the NMSE of SVM+ to the NMSE of SVM

[8] S. Kim, S. Chun, Graded forecasting using an array of bipolar predictions: application of probabilistic neural networks to a stock market index, International Journal of Forecasting, 14 (3), 1998, 323-337.

[9] E. Saad, D. Prokhorov, D. Wunsch, Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks, IEEE Transactions on Neural Networks, 9 (6), 1456-1470, 1998.

[10] Vapnik V. Statistical Learning Theory. Wiley, New York, NY, 1998.

[11] F.E.H. Tay, L.J. Cao, Application of support vector machines in financial time series forecasting, Omega 29 (4), 2001, 309-317.

[12] F.E.H. Tay, L.J. Cao, Improved financial time series forecasting by combining support vector machines with self-organizing feature map, Intell. Data Anal., 5, 2001, 1-16.

[13] Schmidt M. Identifying speaker with support vector networks. Interface 96 Proceedings, Sydney, 1996.

[14] Muller KR, Smola A, Scholkopf B. Prediction time series with support vector machines. Proceedings of International Conference on Artificial Neural Networks, Lausanne, Switzerland, 1997.

[15] Vapnik VN, Golowich SE, Smola AJ. Support vector method for function approximation, regression estimation, and signal processing. Advances in Neural Information Processing Systems, 1996, 281-287.

[16] Refenes, A-P., Zapranis, A.D. and Francis, G. Modeling stock returns in the framework of APT: a comparative study with regression models, Neural Networks in the Capital Markets, 1995, 101-125.

[17] Tsibouris, G. and Zeidenberg, M. Testing the efficient markets hypothesis with gradient descent algorithms, Neural Networks in the Capital Markets, 1995, 127-136.

[18] Steiner, M. and Wittkemper, H-G. Neural networks as an alternative stock market model, Neural Networks in the Capital Markets, 1995, 135-147.

[19] Wittkemper, H-G. and Steiner, M. Using neural networks to forecast the systematic risk of stocks, European Journal of Operational Research, 90, 1996, 577-588.

[20] Shazly, M.R.E. and Shazly, H.E.E. Forecasting currency prices using genetically evolved neural network architecture, International Review of Financial Analysis, 8, No. 1, 1999, 67-82.

[21] Tay FEH, Cao LJ. Improved financial time series forecasting by combining support vector machines with self-organizing feature map. Intelligent Data Analysis, 2001.

[22] Gestel, T.V., Suykens, J.A.K., Baestaens, D-E., Lambrechts, A., Lanckriet, G., Vandaele, B., Moor, B.D. and Vandewalle, J. Financial time-series prediction using least squares support vector machines within the evidence framework, IEEE Transactions on Neural Networks, 12, No. 4, 2001, 809-821.

[23] Wun-Hua Chen and Jen Ying Shih. Comparison of support vector machines and back propagation neural networks in forecasting the six major Asian stock markets, International Journals Electronics Finance, 1, 2006, 49-67.

[24] Chih Fong Tsai. Financial decision support using neural network and support vector machines, Expert Systems, 25, 2008, 380-393.

[25] Tay, F.E.H. and Cao, L. Application of support vector machines in financial time-series forecasting, Omega, 29, 2001, 309-317.

[26] Ince H, Trafalis T. Short term forecasting with support vector machines and application to stock price prediction. International Journal of General Systems, 37(6), 2008, 677-687.

[27] Nayak, R. K., Mishra, D. and Rath, A. K. A Nave SVM-KNN based stock market trend reversal analysis for Indian benchmark indices.. Appl. Soft Comput., 35, 2015, 670-680.

[28] Thomason M. The practitioner methods and tool. Journal of Computational Intelligence in Finance , 7(3), 1999, 36-45.

[29] Juszczak, PD. M. J. Tax; R. P. W. Dui. Feature scaling in support vector data descriptions, Proc. 8th Annu. Conf. Adv. School Comput. Imaging, 2002.

ISSN 2356-5608