International Journal of the Physical Sciences Vol. 6(25), pp. 6091-6097, 23 October, 2011
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS11.466
ISSN 1992 - 1950 ©2011 Academic Journals

Full Length Research Paper

A new support vector machine- genetic algorithm (SVM-GA) based method for stock market forecasting

Vahid Khatibi*, Elham Khatibi and Abdolreza Rasouli

Bardsir Branch, Islamic Azad University, Bardsir, Iran.

Accepted 26 July, 2011

Since analysis of time series is so hard to do, a support vector machine can be more proper for the purpose of forecasting in field of stock market. The support vector machine (SVM) can explore suitable knowledge from so vague data, which usually is necessary to interpret the financial data. But single SVM cannot achieve accurate results. Subsequently, in this paper a combinational intelligent strategy is presented. The proposed strategy consists of genetic algorithm (GA) and SVM for the purpose of stock market forecasting. The genetic algorithm is useful to choose the most informative input indicators from among all the technical indicators. A variety of indicators from the technical analysis field of study are used as input features. Based on obtained results, the hybrid GA-SVM system performs better than Neural Network system.

Key words: Stock price forecast, genetic algorithm, support vector machine.

INTRODUCTION

All the investor needs to know to make a buying or selling decision is the expected direction of the stock. Studies have also shown that predicting direction as compared to value can generate higher profits (Chen et al., 2003). A number of artificial intelligence (AI) and machine learning techniques have been used over the past decade to predict the stock market. Neural networks (NN) are by far the most widely used technique. Time delay neural networks have been used in (Kreesuradej et al., 1994) for stock market trend prediction. Probabilistic neural networks have been used in (Tan et al., 1995) to model it as a classification problem, the 2 classes being a rise or a fall in the market. Recurrent neural nets have been used in (Saad et al., 1996) for predicting the next day's price of the stock index. Other methods that have been used to forecast the stock market include Bayesian forecasting technique (Wolfe, 1988), progressive algorithms (Kanoudan, 2000; Kim, 2000), clustering algorithms (Schulenburg and Ross, 2001) and fuzzy logic (Castillo and Melin, 2001).

Kim and Shin (2007) have proposed a hybrid model of genetic algorithms and neural networks for optimization of the network structure features to present the more accurate results. The study in (Tsaih et al., 1998) combined the rule-based method and ANN to forecast the range of change in the S and P 500 stock indicators based on daily estimates.

Kohara et al. (1997) integrated previous information in ANN to increase the accuracy of stock data estimation. In the last few years, the use of SVMs for stock market forecasting has made significant progress. SVMs were first used by Cao and Tay (2001) and Tay and Cao (2001a, b) for financial time series forecasting.

Kim (2003) has proposed an algorithm to predict the stock market direction by using technical analysis indicators as input to SVMs. Studies have compared SVM with NN and time series techniques. The obtained results proved that proposed method outperformed other used forecasters (Chen and Shih, 2006).

In Henri (2011), the authors have been presented a binary model to estimate the up/down variation for stock in US and in Yakup et al. (2011), the NN algorithms are

*Corresponding author. E-mail: mail:khatibi78@yahoo.com.

Abbreviations: SVM, Support vector machine; **GA,** genetic algorithm; **NN,** neural networks; **EMH,** efficient market hypothesis; **AI,** artificial intelligence; **SXGE,** Swedish stock index; **HR,** hit rate; **RP,** realized potential; **PSO,** particle swarm optimization; **RBF,** radial basis function.

applied to detecting the variation of indexed stock in turkey.

The SVM decreases the level of risk in information data and leads to the more degree of accuracy by using a structural method compare with the NN, which implement the empirical risk minimization principle.

THE PROPOSED HYBRID METHOD

The market action uses past prices and trading volumes to predict the future price. Technical analysis assumes that stock prices move in trends, and that the information which affects prices enters the market over a finite period of time, not instantaneously. Technical analysis contradicts the long held efficient market hypothesis (EMH). EMH expresses that stock price proceeds an uncertain change and cannot be estimated according to the past features. Based on EMH, all information that enters the market affects the prices instantaneously. If the EMH were true, it would not be possible to use AI techniques to predict the market. However, due to the success of technical analysts in the financial world and a number of studies appearing in academic literature successfully using AI techniques to predict the market, EMH is widely believed to be a null hypothesis now.

Technical analysts make use of technical indicators, which are mathematical formulations which give us clues about the trend of the market. An example of a technical indicator is the famous stochastic oscillator $\%K$:

$$\%K = (P(c) - P(l))/(P(h) - P(l)) \quad (1)$$

Where $P(c)$, $P(h)$, and $P(l)$ are closest, highest and lowest price of a security over any time period. Technical analysts normally use a number of such indicators and judgment obtained from deciding which sample a special tool shows at a certain time, and what the description of mentioned sample has to be. These technical indicators have been successfully used as input features to AI techniques, for example, in (Kim, 2003).

The SVMs were introduced by Vapnik (1999). SVMs are a type of maximum margin classifiers. They seek to find a maximum margin bound to set apart the clusters, i.e., they make maximum the difference of the higher bound from the nearest training examples. The higher bound thus obtained is named the optimal higher bound and the training examples that are nearest to the maximum margin higher bound are named support vectors.

If we can map the data in a linear equation, the following equation can interpret which by two features (decision clusters):

$$out = w_0 + w_1 f_1 + w_2 f_2 \quad (2)$$

Where out is the output, f_i are the feature amounts, and there are three weights w_i used by a certain learning algorithm. The maximum margin higher bound is interpreted as the next relation regarding the support vectors:

$$m = c + \Sigma d_i m_i t(i).t \quad (3)$$

where m is the cluster amount of training sample $t(i)$, the array t shows a validation sample, the arrays $t(i)$ are the support vectors and (.) represents the dot product. In this equation, c and d_i are metrics that specify the higher bound. Determining the support vectors and specifying the metrics c and d_i are same as solving a linearly constrained quadratic problem. If mapping the data in a linear equation is impossible, as in this case, SVM transforms the inputs into the high dimension indicators range. A kernel function is used to perform the transformation:

$$y = b + \Sigma \alpha_i y_i K(x(i),x) \quad (4)$$

Many various kernels are used to generate the internal products to build SVM with various models of nonlinear equations in the input range. Most usual kernel functions are Gaussian radial basis function (RBF) and the polynomial function.

$$K(a; b) = exp(-1/\delta^2 (a - b)^2) \quad (5)$$

$$K(a; b) = (ab+1)^n \quad (6)$$

Where n is the degree of the polynomial kernel and δ^2 is the bandwidth of the Gaussian RBF kernel. A unique feature of SVMs is that they are resistant to the over-fitting problem because SVMs rely on structures and regular models compare with the neural networks which try to minimize the risk by using empirical principals. The previous tries to decrease the clustering error or explore from true solution of the training set, but the latest tries to decrease the higher level of generalization error.

The stock market direction problem is modeled as a two class classification problem. The directions are classified as 0 and 1 in the dataset. A class value of 0 indicates that the current day's price is less than the previous day, that is, a fall in the stock, and a class value of 1 indicates that the current day's price is more than the previous day, that is, an increase in the stock price. We chose the Indian stock market for the study.

In the past, most of the work in this area has focused on the American and Korean stock markets; there exists little published work using an AI technique for predicting the Swedish stock market.

This is significant as studies have shown that different stock markets have different characteristics and results obtained for one are not necessarily true for another (Chen and Shih, 2006). Studies have shown that the price of a stock does not move in isolation. There is statistically significant correlation between prices of certain stocks and thus, price movements in one stock can often be used to predict the movement of other stocks (Kim et al., 2002; Kwon et al., 2005).

Let the two stocks whose correlation we want to find be denoted by P and Q. The correlation between these stocks is given by:

$$Cor(P,Q) = \Sigma ((P(i) - QA)(P(i) - QA))/(\sigma_S \sigma_T n) \quad (7)$$

Where $P(i)$ and $Q(i)$ are stock prices on the ith day, PA and QA are the mean prices of the stocks, σ_S and σ_T are the standard deviations, and n is the number of days over which the correlation is to be found.

Technical analysts make use of technical indicators, which are mathematical formulations which give us clues about the trend of the market. We use a set of 35 such technical indicators as candidates for input features that are being used by financial experts (Kaufman, 1998). Some of the more important features are given in Table 1.

We first find the m companies which exhibit the highest correlation with the stock to be predicted. One of these m stocks will always be the target stock itself as it will have perfect correlation with itself. Then, these 35 features are calculated for each of these m companies by using their past prices and trading volumes. Thus, we obtain a set of 35*m candidate features. As explained above, we obtain a set of 35*m candidate features. Now GA is applied to choose a set of most important indicators from among them. The selected features are used as inputs to a SVM.

The purpose here is to gain the best subset of indicators which produce the most accurate results. The various steps in the GA are described below:

- Representation: A chromosome is presented with a binary array of size 35*m, that each bit in the chromosome indicates whether the corresponding feature is selected.
- Fitness Evaluation: The following fitness function is used for

Table 1. Some features and their formulas.

Feature name	Formula
Momentum	(C(i)/C(i-N))* 100
Williams %R	(HH(n)-C(t))/ (HH(n)-LL(n))*100
Rate of Change (ROC)	(C(t) – C(t-n))/C(t-n)
5 Day Disparity	(C(t)/MA(5))*100
10 Day Disparity	(C(t)/MA(10))*100
Stochastic %K	(C(t) – L(t))/(H(t) – L(t))
Price Volume Trend (PVT)	((C (t) – C (t-1)) / C(t-1)) * v

Table 2. Forecasting in SXGE.

Period	$HR_{Proposed}$	$RP_{Proposed}$	HR_{naive}	RP_{naive}
1993	52.88	12.96	54.7	22.3
1994	61.93	21.36	52	15.2
1995	59.95	13.2	51.4	8.9
1996	63.69	29.4	50	1.7
1997	65.01	38.88	52.8	12.5
1998	63.69	29.28	53.5	18.3
1999	66.11	31.32	53.9	12
2000	61.82	25.2	48.5	4.3
Average	**61.885**	**25.2**	**52.1**	**11.9**

evaluating the fitness of a chromosome i:

$$\text{Fitness} = (A(i) - AR)/(\Sigma (A(i) - AR)) \quad (8)$$

Where, $A(i)$ is the classification accuracy obtained by the SVM with the input feature set as described by chromosome I and AR is the accuracy of a random guess, which, in this case is 0.5.

- Selection: Parent selection is performed by Roulette Wheel selection. Thus, chromosomes with high fitness scores get selected more often.
- Crossover and Mutation are then carried out to produce a new generation.
- Stopping Condition: If finding the better solution for a determined number of generations is not possible the GA stops.

The optimal set of features as selected by the genetic algorithm above is then used as input to the SVM. The original input features are converted into the interval of [–1, 1]. The aim of linear mapping is to freely standardize each indicator to the determined interval. It guarantees that the big amount input features do not defeat small amount inputs, and thus forecasting errors will be reduced.

The SVM Light software package was used to perform the experiment. The kernel function used for transforming the input space to the higher dimension space is the Gaussian radial basis function kernel. This kernel function was selected as it gave better experimental results than the other common kernel functions.

PERFORMANCE CRITERIA

Predictions in this filed usually are evaluated by three major metrics; Realized Potential (R_P), Return On Investment (ROI) and Hit Rate (H_R).

Hit rate is formulized as below;

$$H_R = \frac{\left|\left\{i \mid \Lambda_i^k \times \overline{\Lambda}_i^k > 0, i = 1...N\right\}\right|}{\left|\left\{i \mid \Lambda_i^k \times \overline{\Lambda}_i^k \neq 0, i = 1...N\right\}\right|} \quad (9)$$

Where Λ_i^k and $\overline{\Lambda}_i^k$ denote the actual and forecasted value in ith time at iteration k, respectively.

Return on investment indicates the effects of predictions on total returns over the period of T.

$$ROI = \sum_{i=1}^{N} \Lambda_i . sign(\overline{\Lambda}_i) \quad (10)$$

And realized potential i computed as follows;

$$RP = \frac{\sum_{i=1}^{N} \Lambda_i . sign(\overline{\Lambda}_i)}{\sum_{i=1}^{N} |\Lambda_i|} \quad (11)$$

All forecasting methods in this area use the naive predictor to evaluate the performance of the prediction. The naive predictor is the standard trivial forecaster which forecasts the future investment according to the current price. It is necessary that the proposed prediction methods outperform the naive method.

NEUMERICAL RESULTS

In this section, results of using the proposed method are presented. We are going to perform the prediction related to the three datasets comprising stock market prices in eight years; Swedish stock index (SXGE), Ericsson and Volvo. Since all stock price forecasters should present the better results compared with the naive strategy, all comparisons in this section are organized based on proposed method results versus the naive strategy as described in previous section.

Also, all performance parameters described in the previous section are computed. For each dataset the eight years between 1993 and 2000 are considered as the period of prediction. This wide range could be useful to reveal the real performance of proposed method. Accessing to newer data is impossible due to some security problems. Nevertheless, the existence datasets are enough to evaluation of proposed method.

Hit Rate (H_R) and Realized potential (R_P) parameters are computed for proposed and naive method in Table 2, 3 and 4. Tables show the mentioned parameters for SXGE, Ericsson and Volvo datasets respectively. The mean of all computed H_R and R_P based on naive and proposed method is the main parameter to comparison.

In addition, the progress trend of ROI according to the specified time periods (Eight Periods) is presented in Figure 1, 2 and 3. ROI values represented in the figures

Table 3. Forecasting in Ericsson.

Period	$HR_{Proposed}$	$RP_{Proposed}$	HR_{naive}	RP_{naive}
1993	47.56	12.12	48	22.6
1994	56.32	11.16	41.7	2.1
1995	61.27	21.36	48.2	4.5
1996	56.32	18.24	39.8	-6.4
1997	61.93	36.96	43.3	-7.3
1998	52.8	0	46.5	-0.4
1999	58.85	4.44	40.7	-0.5
2000	55.77	6.12	42.4	-5.8
Average	**56.3525**	**13.8**	**43.825**	**1.1**

Table 4. Forecasting in Volvo.

Period	$HR_{Proposed}$	$RP_{Proposed}$	HR_{naive}	RP_{naive}
1993	53.88	11.64	41.3	12.8
1994	57.64	21.2	45.7	22.3
1995	44.33	-3	37.9	1.5
1996	51.15	16.56	39	5.4
1997	48.95	1.56	42.5	5.4
1998	47.51	-3	47.6	18.4
1999	49.39	9.36	42.9	10.6
2000	52.58	18	45.8	12.2
Average	**50.67875**	**9.04**	**42.8375**	**11.075**

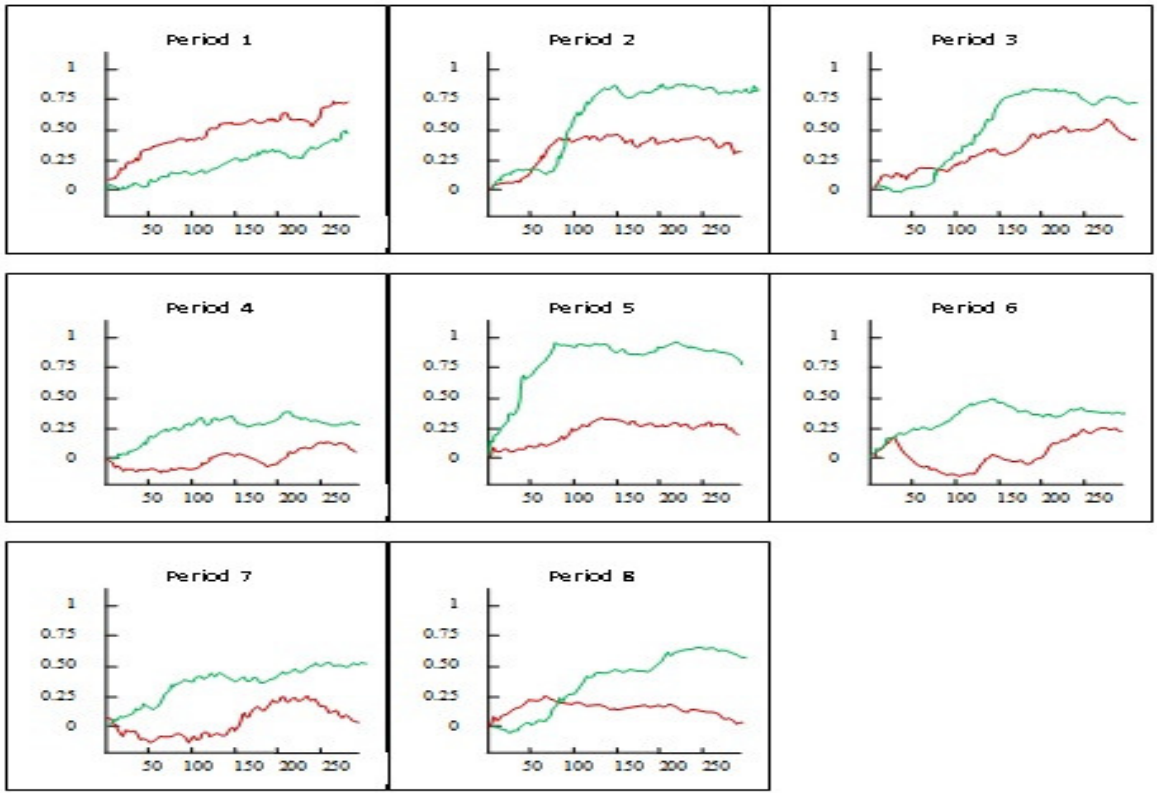

Figure 1. ROI for SXGE.

Table 5. Forecasting in Volvo against PSO.

Period	$HR_{Proposed}$	$RP_{Proposed}$	HR_{PSO}	RP_{PSO}
1993	53.88	11.64	44.1	13.8
1994	57.64	21.2	45.9	21.3
1995	44.33	-3	34.9	1.5
1996	51.15	16.56	39	5.7
1997	48.95	1.56	42.5	5.1
1998	47.51	-3	47.6	18.4
1999	49.39	9.36	44.9	12.6
2000	52.58	18	45.8	12.2
Average	**50.67875**	**9.04**	**42.2875**	**11.325**

Table 6. Forecasting in Volvo Vs MLP.

Period	$HR_{Proposed}$	$RP_{Proposed}$	HR_{MLP}	RP_{MLP}
1993	53.88	11.64	49.6	12.9
1994	57.64	21.2	44.4	22.4
1995	44.33	-3	35.9	1.6
1996	51.15	16.56	34	5.7
1997	48.95	1.56	44.5	5.1
1998	47.51	-3	48.6	19.1
1999	49.39	9.36	41.1	13.2
2000	52.58	18	44.5	13.6
Average	50.67875	9.04	42.8875	11.825

have been computed daily.

Green line and red line show the proposed and naive methods respectively. All holidays have been excluding from each period. For more comparison the proposed method has been compared with Particle Swarm Optimization (PSO) and MLP on VOLVO data, as seen in Tables 5 and 6.

DISCUSSION

Based on the results presented in Table 2, 3 and 4 the proposed method outperforms in many periods. All periods in SXGE show the noteworthy improvement in both H_R and R_P excluding period 1. The most significant enhancement according to the amount of H_R and R_P is related to the Ericsson dataset. In Volvo dataset a noticeable improvement is seen in term of H_R but improvement of Rp is not so considerable.

On the other hand, Figures 1, 2 and 3 depict the improvement of Return on Investment (ROI) by using the proposed method. Period 5 in SXGE, period 5 in Ericsson and period 7 in Volvo are the best results regarding the ROI in each dataset. It should be noticed that in some periods the proposed method produces the inaccurate estimates, for example period 1 in SXGE, period 1 in Ericsson and finally period 5, 6 in Volvo. To sum up, as can be seen, the most significant improvement has been appeared by performing the forecast in SXGE dataset and the lowest improvement is related to Volvo dataset.

According to the above discussion, it is proved that the proposed hybrid system (SVM-GA) can forecast the future price in stock markets with reasonable error rate.

CONCLUSION

In this paper, we proposed a hybrid GA-SVM system for predicting the future direction of stock prices. A set of technical indicators, obtained from the stock to be predicted, and also from the stocks exhibiting high correlation with that stock were used as input features. The results showed that the correlation concept and the GA helped in improving the performance of the SVM system significantly. There is a lot of scope for further work in this area. If various political and economic factors which affect the stock market are also taken into consideration other than the technical indicators as input variables, better results may be obtained. Also, incorporating market specific domain knowledge into the system might help in achieving better performance.

ACKNOWLEDGMENTS

The author wish to express his sincere gratitude to

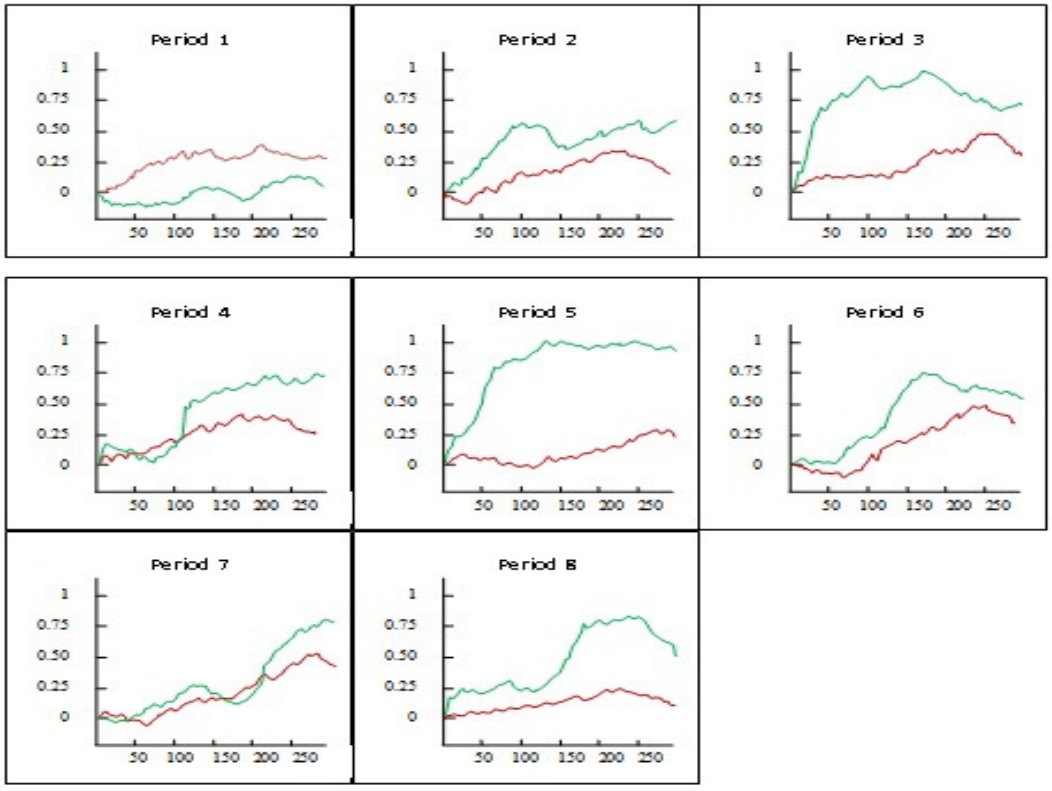

Figure 2. ROI for Ericsson.

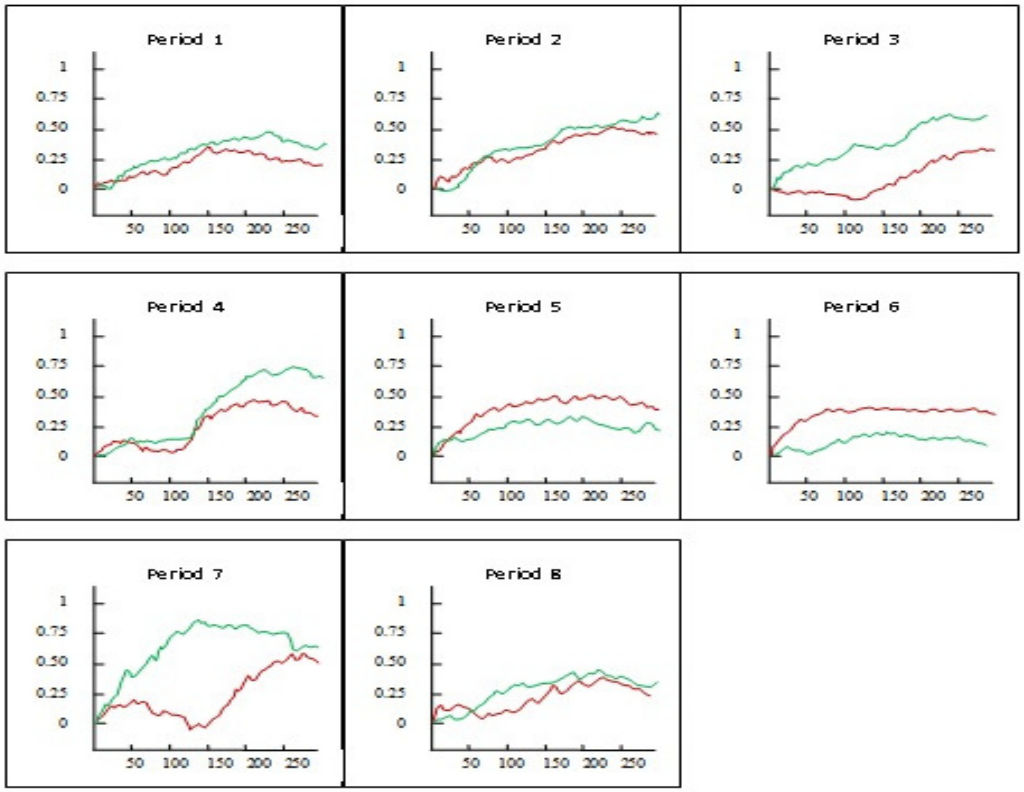

Figure 3. ROI for Volvo.

Islamic Azad University – Bardsir branch, Research Management Center for providing me an opportunity to do my project works on “Proposing a new intelligent method for stock market price”. This project bears on imprint of many peoples.

REFERENCES

Castillo O, Melin P (2001). Simulation and forecasting complex financial time series using neural networks and fuzzy logic, Proceedings of IEEE Conference on Systems, Man, Cybernetics, pp. 2664-2669.

Cao LJ, Tay FEH (2001). Financial forecasting using support vector machines, Neural Comput. Appl., 10: 184-192.

Chen AS, Leung MT, Daouk H (2003). Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index. Comput. Operations Res., 30: 901-923.

Chen WH, Shih JY (2006). Comparison of support-vector machines and back propagation neural networks in forecasting the six major Asian stock markets, Int. J. Elect. Finance, 1: 1.

Henri N (2011). Forecasting the direction of the US stock market with dynamic binary probit models, Int. J. Forecasting, 27(2): 561-578.

Kaufman PJ (1998). Trading Systems and Methods, John Wiley & Sons.

Kanoudan MA (2000). Genetic programming prediction of stock prices. Comput. Econ., 16: 207-236.

Kim HJ, Lee Y K, Kahng BN, Kim IM (2002). Weighted scale-free network in financial correlation, J. Phys. Soc. Japan, 71(9): 2133-2136.

Kim H, Shin K (2007). A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets, Appl. Soft Comput., 7(2): 569-576.

Kim K (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert Syst. Appl., 19(2): 125-132.

Kim K (2003). Financial time series forecasting using Support Vector Machines, Neuro comput., 55: 307-319.

Kreesuradej W, Wunsch D, Lane M (1994). Time-delay neural network for small time series data sets, in World Cong. Neural Networks, SanDiego, CA.

Kohara K, Ishikawa T, Fukuhara Y, Nakamura Y (1997). Stock price prediction using prior knowledge and neural networks. Int. J. Intell. Syst. Accounting, Finance Manag. 61: 11–22.

Kwon YK, Choi SS, Moon BR (2005). Stock prediction based on financial correlation, GECCO, pp. 2061-2066.

Saad E, Prokhorov D, Wunsch D (1996). Advanced neural-network training methods for low false alarm stock trend prediction, in Proc. IEEE Int. Conf. Neural Networks, Washington, D.C.

Schulenburg S, Ross P (2001). Explorations in LCS models of stock trading, Advances in Learning Classifier Systems, pp. 151-180.

Tan H, Prokhorov D, Wunsch D (1995). Probabilistic and time-delay neural-network techniques for conservative short-term stock trend prediction, in Proc. World Congr. Neural Networks, Washington, D.C.

Tsaih R, Hsu Y, Lai CC (1998). Forecasting S&P 500 stock index futures with a hybrid AI system. Decis. Support Syst., 23(2): 161-174.

Tay FEH, Cao LJ (2001a). Application of support vector machines in financial time series forecasting. Omega, 29: 309-317.

Tay FEH, Cao LJ (2001b). Improved financial time series forecasting by combining support vector machines with self-organizing feature map. Intell. Data Anal., 5: 339-354.

Vapnik VN (1999). An overview of statistical learning theory. IEEE Trans Neural Networks, 10: 988-999.

Wolfe RK (1988). Turning point identification and Bayesian forecasting of a volatile time series, Comput. Industrial Eng., 378-386.

Yakup K, Melek AB, Ömer KB (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange, Expert Syst. Appl., 38(5): 5311-5319