World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:3, 2008

A Hybrid Machine Learning System for Stock Market Forecasting

Rohit Choudhry, and Kumkum Garg

***Abstract*—**In this paper, we propose a hybrid machine learning system based on Genetic Algorithm (GA) and Support Vector Machines (SVM) for stock market prediction. A variety of indicators from the technical analysis field of study are used as input features. We also make use of the correlation between stock prices of different companies to forecast the price of a stock, making use of technical indicators of highly correlated stocks, not only the stock to be predicted. The genetic algorithm is used to select the set of most informative input features from among all the technical indicators. The results show that the hybrid GA-SVM system outperforms the stand alone SVM system.

***Keywords*—**Genetic Algorithms, Support Vector Machines, Stock Market Forecasting.

I. Introduction

STOCK market prediction is regarded as a challenging task in financial time-series forecasting. This is primarily because of the uncertainties involved in the movement of the market. Many factors interact in the stock market including political events, general economic conditions, and traders' expectations. Therefore, predicting market price movements is quite difficult. Increasingly, according to academic investigations, movements in market prices are not random. Rather, they behave in a highly non-linear, dynamic manner. Also, the ability to predict the direction and not the exact value of the future stock prices is the most important factor in making money using financial prediction. 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 [1].

The rest of this paper is organized as follows: In section 2, we give an overview of previous studies in this area. In sections 3 and 4, we give a brief introduction to the basic concepts behind the theory of technical analysis and SVM respectively. In section 5, the stock prediction problem is explained. In section 6, we describe our proposed system. In section 7, the experimental results are given. Conclusions and directions for further work are given in section 8.

Manuscript received March 29, 2008. Rohit Choudhry is a Masters' student at the Electronics & Computer Engineering Department, Indian Institute of Technology Roorkee, India (e-mail: rohetuec@iitr.ernet.in).

Kumkum Garg is Professor at the Electronics & Computer Engineering Department, Indian Institute of Technology Roorkee, India (e-mail: kgargfec@iitr.ernet.in).

II. Related Research

A number of artificial intelligence and machine learning techniques have been used over the past decade to predict the stock market. Neural Networks are by far the most widely used technique. Time Delay Neural Networks have been used in [2] for stock market trend prediction. Probabilistic Neural Networks have been used in [3] 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 [4] for predicting the next day's price of the stock index. Other methods that have been used to forecast the stock market include Bayesian belief networks [5], evolutionary algorithms [6] [7], classifier systems [8], and fuzzy sets [9].

Recent research tends to hybridize other AI techniques with ANN. Kim & Shin [10] have proposed a hybrid model of Genetic Algorithms and Neural Networks for optimization of the number of time delays and network architectural factors using GA, to improve the effectiveness of constructing the ANN model. The study in [11] integrated the rule-based technique and ANN to predict the direction of change of the S&P 500 stock index futures on a daily basis. Kohara et al [12] incorporated prior knowledge in ANN to improve the performance of stock market prediction.

In the last few years, the use of SVMs for stock market forecasting has made significant progress. SVMs were first used by Tay & Cao for financial time series forecasting [13], [14], [15]. Kim has proposed an algorithm to predict the stock market direction by using technical analysis indicators as input to SVMs [16]. Studies have compared SVM with Back Propagation Neural Networks (BPN). The experimental results showed that SVM outperformed BPN most often though there are some markets for which BPN have been found to be better [17]. These results may be attributable to the fact that the SVM implements the structural risk minimization principle and this leads to better generalization than Neural Networks, which implement the empirical risk minimization principle.

III. Technical Analysis

Technical analysis is the study of market action using past prices and trading volumes for the purpose of forecasting future price trends. 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 states that market

prices follow a random walk and cannot be predicted based on their past behavior. According to 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:

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

where P(c), P (h), and P (l) represent closing price, highest price and lowest price of a security over any time period. Technical analysts normally use a number of such indicators and judgment gained from experience to decide which pattern a particular instrument reflects at a given time, and what the interpretation of that pattern should be. Technical analysts may disagree among themselves over the interpretation of a given chart. These technical indicators have been successfully used as input features to AI techniques, for example, in [16].

IV. Support Vector Machines

The Support Vector Machines (SVMs) were proposed by Vapnik [18]. SVMs are a type of maximum margin classifiers. They seek to find a maximum margin hyperplane to separate the classes, i.e., they maximize the distance of the hyperplane from the nearest training examples. The hyperplane thus obtained is called the optimal separating hyperplane (OSH) and the training examples that are closest to the maximum margin hyperplane are called support vectors.

If the data is linearly separable, a hyperplane separating the binary decision classes in the two attribute case can be represented as the following equation:

$$y = w_0 + w_1x_1 + w_2x_2 \quad (1)$$

where y is the outcome, x_i are the attribute values, and there are three weights w_i to be learned by the learning algorithm. The maximum margin hyperplane can be represented as the following equation in terms of the support vectors:

$$y = b + \sum \alpha_i y_i x(i).x \quad (2)$$

where y is the class value of training example x(i), the vector x represents a test example, the vectors x(i) are the support vectors and · represents the dot product.. In this equation, b and α_i are parameters that determine the hyperplane. Finding the support vectors and determining the parameters b and α_i is equivalent to solving a linearly constrained quadratic programming problem.

If the data is not linearly separable, as in this case, SVM transforms the inputs into the high-dimensional feature space. This is done by using a kernel function as follows:

$$y = b + \sum \alpha_i y_i K(x(i),x) \quad (3)$$

There are many different kernels for generating the inner products to construct machines with different types of nonlinear decision surfaces in the input space. Common examples of the kernel function are the polynomial kernel $K(x; y) = (xy+1)^n$ and the Gaussian radial basis function(RBF) $K(x; y) = \exp(-1/\delta^2(x - y)^2)$ 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. This is because while many traditional neural network models have implemented the empirical risk minimization principle, SVM implements the structural risk minimization principle. The former seeks to minimize the misclassification error or deviation from correct solution of the training data, but the latter searches to minimize an upper bound of generalization error.

V. The Stock Direction Prediction Problem

The stock market direction problem is modeled as a two class classification problem. The directions are categorized as 0 & 1 in the data. A class value of 0 means that the present day's price is less than the previous day, i.e., a fall in the stock, and a class value of 1 means that the present day's price is more than the previous day, i.e., a rise 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 Indian 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 [17]. In the Indian stock market, we have chosen 3 stocks; Tata consultancy services (TCS), Infosys, and Reliance industries limited (RIL) for our experiments as these are the most prominent stocks on India's stock exchange.

VI. Proposed System

A. Correlation between Stocks

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 [19] [20].

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

$$Cor(S, T) = \sum ((S(i) - SA)(T(i) - TA))/(\sigma_S \sigma_T n)$$

where S(i) & T(i) are closing prices of the stock on the ith day, SA & TA 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.

As an example, Fig. 1 shows five companies having the highest absolute value of correlation with TCS, which is a major IT services provider and a part of the Tata

conglomerate. It can be seen that the companies having the highest level of correlation with TCS are the ones which are in the same industry or a part of the same group; a result that was expected.

Tata Consultancy Services – An IT services firm

Highly Correlated Companies:

- Infosys Technologies – An IT services firm
- Wipro Technologies – An IT services firm
- Tata Motors – A motor company, also part of the Tata group
- Bharti – India's largest telecom company

Fig. 1 Correlation example of a company

B. Input Features

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 [21]. Some of the more important features are given in Table I.

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.

C. Genetic Algorithm

As explained above, we obtain a set of 35*m candidate features. A Genetic Algorithm is now used to select a set of salient features from among them. The selected features are used as inputs to a Support Vector Machine. The purpose here is to obtain an optimal subset of features which produce the best possible results. The various steps in the GA are described below:

- Representation: We represent a chromosome by a binary vector of size 35*m, where each bit of the chromosome tells whether the corresponding input feature is selected or not.
- Fitness Evaluation: The following fitness function is used for evaluating the fitness of a chromosome i:

$$\text{fitness} = (A(i) - A_R)/(\sum (A(i) - A_R))$$

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

- Selection: Roulette Wheel selection is used for parent 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: The GA stops when it does not find a better solution for a fixed number of generations.

D. Support Vector Machine

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 scaled into the range of [−1,1]. The goal of linear scaling is to independently normalize each feature component to the specified range. It ensures the larger value input attributes do not overwhelm smaller value inputs, and thus helps to reduce prediction errors. The SVM Light software package [22] 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.

TABLE I
SOME OF THE INPUT 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

VII. EXPERIMENTAL RESULTS

We tested our approach with three stocks, TCS, Infosys and RIL as mentioned above. The data used for this study were obtained from the Yahoo Finance website [23]. We collected in all 1386 trading days' data from August 12, 2002 to January 18, 2008. For each day, the opening, highest, lowest and closing values of the stock price were obtained. Further, the trading volumes were also obtained. The data were collected for the 30 companies which comprise the Bombay Stock Exchange's representative index 'Sensex'. 60% of the data was used for training, 20% for validation and 20% for testing the system.

The prediction performance is measured in terms of 'hit ratio', which is the percentage of times our system's prediction for direction was correct. The results of our approach were compared with the results obtained by the stand alone SVM, where the 35 features for the target

company alone are used as input. The results of the stand alone SVM were found to match very closely with the results reported by Kim [16] for a similar model. Also, our GA-SVM hybrid model significantly outperformed the SVM. For example, for TCS, the hit ratio of our GA-SVM was 61.7328% while that of the SVM was found to be 58.0903%. The hit ratios for all the three stocks are given in Table II.

TABLE II
HIT RATIOS OF SVM AND GA-SVM (%)

	SVM	**GA-SVM**
TCS	58.09	61.732
Infosys	56.748	60.285
Reliance	55.643	59.534

VIII. 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 & 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 & 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.

REFERENCES

[1] Chen, A.S., Leung, M.T., and Daouk, H. Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index. Computers and Operations Research 30, 2003, 901-923.
[2] W. Kreesuradej, D. Wunsch, and M. Lane, Time-delay neural network for small time series data sets, in *World Cong. Neural Networks,* San Diego, CA, June 1994.
[3] H. Tan, D. Prokhorov, and D. Wunsch, Probabilistic and time-delay neural-network techniques for conservative short-term stock trend prediction, in *Proc. World Congr. Neural Networks,* Washington, D.C., July 1995.
[4] E. Saad, D. Prokhorov, and D. Wunsch, Advanced neural-network training methods for low false alarm stock trend prediction, in *Proc. IEEE Int. Conf. Neural Networks,* Washington, D.C., June 1996.
[5] R. K. Wolfe, Turning point identification and Bayesian forecasting of a volatile time series, *Computers and Industrial Engineering*, 1988, pp 378–386.
[6] M. A. Kanoudan, Genetic programming prediction of stock prices. *Computational Economics*, 16, 2000, pp 207–236.
[7] K. J. Kim. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications*, 19(2), 2000, pp 125–132.
[8] S. Schulenburg and P. Ross, Explorations in LCS models of stock trading, *Advances in Learning Classifier Systems*, 2001, pages 151–180.
[9] O. Castillo and P. Melin, Simulation and forecasting complex financial time series using neural networks and fuzzy logic, *Proceedings of IEEE Conference on Systems, Man, and Cybernetics*, 2001, pages 2664–2669.
[10] H Kim and K Shin, A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets, Applied Soft Computing, Volume 7, Issue 2, March 2007, Pages 569-576.
[11] Tsaih, R., Hsu, Y. and Lai, C.C., Forecasting S&P 500 stock index futures with a hybrid AI system. *Decision Support Systems* 23 2, 1998, pp. 161–174.
[12] Kohara, K., Ishikawa, T., Fukuhara, Y. and Nakamura, Y., Stock price prediction using prior knowledge and neural networks. *International Journal of Intelligent Systems in Accounting, Finance and Management* 6 1, 1997, pp. 11–22.
[13] L.J. Cao and F.E.H. Tay, Financial forecasting using support vector machines, *Neural Computing Applications* 10, 2001, pp. 184–192.
[14] F.E.H. Tay and L.J. Cao, Application of support vector machines in financial time series forecasting. *Omega* 29, 2001, pp. 309–317.
[15] F.E.H. Tay and L.J. Cao, Improved financial time series forecasting by combining support vector machines with self-organizing feature map. *Intelligent Data Analysis* 5, 2001, pp. 339–354.
[16] K Kim, Financial time series forecasting using Support Vector Machines, Neurocomputing 55, May 2003, Pages 307 – 319.
[17] 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, *Int. J. Electronic Finance, Vol. 1, No. 1, 2006.*
[18] V.N. Vapnik, An overview of statistical learning theory. *IEEE Transactions of Neural Networks* 10, 1999, pp. 988–999.
[19] H. J. Kim, Y. K. Lee, B. N. Kahng, and I. M. Kim, Weighted scale-free network in financial correlation, *Journal of the Physical Society of Japan*, 71(9), 2002, pp 2133–2136.
[20] Y. K. Kwon, S. S. Choi, B. R. Moon, Stock prediction based on financial correlation, GECCO, 2005, pp 2061-2066.
[21] P. J. Kaufman, Trading Systems and Methods, John Wiley & Sons, 1998.
[22] http://svmlight.joachims.org/
[23] http://in.finance.yahoo.com/