A GA-optimized SAX- ANN based Stock Level Prediction System

Binoy B. Nair
(Corresponding Author)
Amrita School of Engineering
Coimbatore Campus
Amrita Vishwa Vidyapeetham
(University)

Nikhil Xavier
Amrita School of Engineering
Coimbatore Campus
Amrita Vishwa Vidyapeetham
(University)

V.P Mohandas, Adarsh Sathyapal, Anusree E G, Pawan Kumar, Vignesh Ravikumar
Amrita School of Engineering
Coimbatore Campus
Amrita Vishwa Vidyapeetham
(University)

ABSTRACT
Forecasting stock price movements is of immense importance to any stock trader. However, traditionally, this has been accomplished using technical analysis tools. In this study, an attempt has been made to employ data mining to identify the one-day-ahead stock price levels. Two different approaches are considered. The two approaches are empirically validated on twelve stock price datasets, with the stocks drawn from the Indian, US and UK stock markets. Results indicate that both the approaches proposed in the present study are capable of successfully forecasting the one-day-ahead stock price levels.

General Terms
Pattern Recognition, stock level prediction

Keywords
Stock, prediction, Artificial Neural Network, Genetic Algorithms, Wavelet, Symbolic Aggregate Approximation

1. INTRODUCTION
Investing in stocks is generally considered to be a high-risk enterprise, mainly due to the apparent unpredictability in the stock price movements. It was claimed in [1] and [2] that stock prices follow random walk and concluded that it is not possible to successfully predict stock prices. However, there is empirical evidence, eg. in [3],[4]and [5]to suggest that it is in fact possible to successfully forecast stock price movements. The traditional approach to forecasting stock markets is to employ technical analysis [6]. However, soft computing based approaches are now being increasingly used for the purpose. Two common approaches to forecasting stock markets using soft computing techniques appear to be (a) forecasting the future stock price values and (b) forecasting the future stock price trends. Hybrid adaptive filters have been used in [7] to forecast stock indices. An empirical study of the effectiveness different types of artificial neural networks (ANN) in predicting the BSE-Sensex was carried out in [8]. It was observed in [8] that Levenberg-Marquardt (LM) algorithm is best suited to train a feed forward ANN. Hence, in the present study, LM algorithm has been used. Neuro-fuzzy hybrid systems and Genetic Algorithm (GA)-ANN hybrid systems have also been used in [9] and [10], respectively, for stock market index prediction. Stock market trend prediction was successfully carried out in [11]-[14]. While a decision tree—rough set hybrid system was used in [11], a hybrid decision tree-neuro-fuzzy System was employed in [12]. In [13], a GA optimized decision tree-Support Vector Machine (SVM) based stock market trend prediction system was used and in [14], the effectiveness of ant colony based forecasting systems in forecasting trends in BSE-Sensex was empirically evaluated. Other approaches considered include temporal association rule mining [15]. Soft computing techniques have also been used for portfolio optimization [16].

In this study, two soft computing based approaches to forecasting one-day-ahead stock price levels are empirically evaluated. The first approach views the stock price level forecasting problem as a regression problem and attempts to forecast one-day ahead stock price trends using a GA optimized symbolic aggregate approximation (SAX)-ANN based regression system. The second approach views the level forecasting problem as a classification problem and uses a similar GA-optimized-SAX-ANN based system but the system is now configured as a classifier. The detailed working of the systems proposed is described in the following sections. The rest of the paper is organized as follows: Section 2 presents the system description, section 3 presents the system specifications. Results are presented in section 3 and conclusions in section 4.

2. SYSTEM DESCRIPTION
Two different approaches were considered for forecasting the one-day-ahead price levels: regression and classification.

The steps involved are detailed in sections below.

2.1 Identification of Chaos
Stock market data tends to exhibit nonlinear dynamics in time domain. This nonlinearity can lead to poor prediction accuracy and hence poor trading performance. In the present study, as the first step, the stock price time series under consideration are tested for chaos using the technique described in [18] and [19].

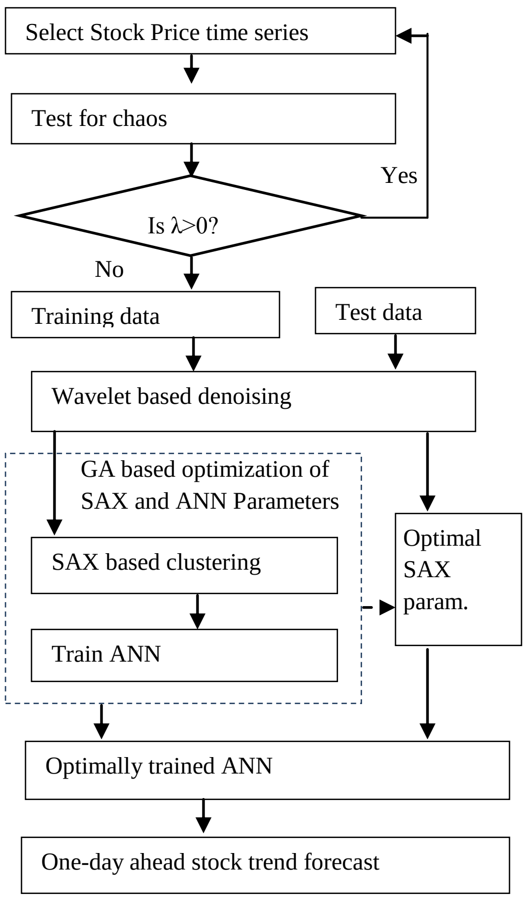


Fig. 1 Proposed system block diagram

The degree of chaos in the selected time series is measured by identifying the value of its Lyapunov exponent (λ). In general, it can be said that:

If $\lambda < 0$; The time series is likely to show deterministic behavior and hence higher predictability .

If $\lambda > 0$; The time series is Chaotic.

If $\lambda = 0$; The time series is in steady state mode and is also likely to exhibit deterministic behavior.

Table 2 shows the corresponding Lambda (λ) values for the selected stocks.

2.2 Wavelet based Denoising

The initial input feature set is represented by the stock price time series vector **X** and the initial output feature set by the one-day-ahead stock price vector **Y.**

$$\boldsymbol{X} = \{y_{t-1}, y_{t-2}, y_{t-3}, \dots, y_{t-N}\}$$
$$\boldsymbol{Y} = \{y_t, y_{t-1}, y_{t-2}, \dots, y_{t-(N-1)}\}$$

The selected stock price datasets are then subjected to denoising. Moving averages are considered to be the simplest type of denoising technique. However, the smoothed waveform is a delayed version of the original time series. The simple Haar Wavelet based denoising is carried out. The Haar wavelet equation is given by:

$$C_i = \frac{y_i - y_{i-1}}{2}$$

Where C_i is the wavelet coefficient and the Haar wavelet scaling function is given by

$$a_i = \frac{y_i + y_{i-1}}{2}$$

Where a_i is the smoothed value.

Three level decomposition is carried out. Reconstruction is then carried out using the technique described in [20] to remove noise.

2.3 SAX based Transformation

Once the waveform reconstruction is performed, SAX algorithm [17] is then used to transform the target vector ***Y*** into its symbol equivalent. As an example, considering the alphabet size to be *d*, the set of symbols is $\boldsymbol{S} = \{S_1, S_2, .., S_d\}$. Each unique symbol now represents a distinct price level. The range of values *y* for each symbol is identified by SAX based on the assumption of the data being normally distributed. Let the length of sliding window be 1, then, the transformed symbol vector will have *N* samples, each of which is represented using one of the *d* symbols as:

$$\boldsymbol{Y_{symbol}} = \{S_{1_t}, S_{2_{t-1}}, S_{d_{t-2}}, \dots, S_{3_{t-N}}\} \text{ etc.}$$

The decision table that can now be used for forecasting is represented as $\{\boldsymbol{X}, \boldsymbol{Y_{symbol}}\}$. It must be noted that in this study, both the alphabet size in SAX and the number of hidden neurons for the ANN are optimized using GA.

2.4 ANN based Forecasting Systems

ANN is used in the present study to forecast the one-day-ahead stock price level. The ANN used in the present study is a single hidden layer, feedforward network trained using Levenberg-Marquardt (LM) algorithm. The LM algorithm was chosen due to its superior performance as demonstrated in [8]. The two approaches proposed in the present study vary in the way the ANN is trained. In the regression based approach, a $1\times h\times 1$ ANN is trained to fit a curve that best forecasts the one-day-ahead price level. In the second approach, a $1\times h\times d$ ANN is trained to classify the forecast into one of the *d* price levels. The number of hidden neurons is optimized using GA. For the ANN, the minimum performance gradient chosen was 10^{-6} and the Max. no. of epochs were chosen to be 100.

2.5 Genetic Algorithm (GA)

GA is widely used for optimization. In this study, GA is used to find the optimal SAX alphabet size and the optimal number of hidden neurons. The GA parameters used in the present study are:

Population size: 20

Max. Generations: 100

Termination condition: 30 generations without improvement in fitness function

Selection: Roulette Wheel selection

Reproduction: Elitism

Mutation: Uniform mutation. Mutation rate= 0.01

3. IMPLEMENTATION RESULTS

The proposed system was empirically validated on twelve different stock price datasets. The stocks considered have been taken from Indian, UK and US stock markets. The Indian stock market considered was the National Stock Exchange (NSE), the UK stock market considered was the London Stock Exchange (LSE) and the US stock markets considered were New York Stock Exchange (NYSE) and NASDAQ.

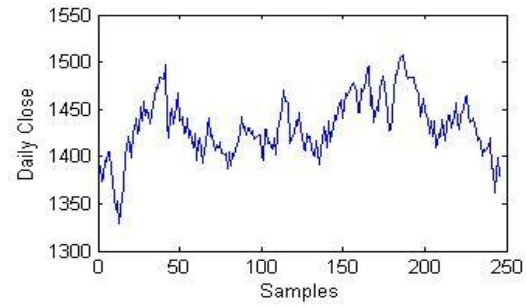
The time frame considered was randomly chosen to be from November 8, 2011 to November 7, 2012. Initial 90% of the data was used for training and the remaining 10% was used for testing. The exact number of samples selected using the above criterion is presented in Table 1.

Table 1. Number of samples

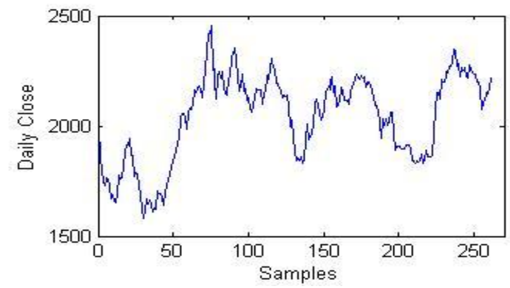
Stock	Training	Testing
	India	
NTPC	234	27
ICICI	216	24
SBI	234	26
TCS	216	24
	US	
Citibank	225	25
Google	225	25
Microsoft	225	25
Royal Dutch Shell	225	25
	UK	
GSK	221	24
RBS	221	25
Rolls Royce Holding	220	24
Sainsbury	234	26

The datasets selected are also representative of the common trends stock movements, eg. uptrend, downtrend and no trend. Figure 2 represents the GSK, SBI and SBY stock price datasets. It can be seen from Figure 2 that while GSK stock price dataset exhibits an initial no-trend pattern followed by a downtrend in later stages, the SBY stock price data exhibits an initial no-trend pattern followed by distinct uptrend. SBI stock price data is seen to exhibit cyclic behavior (no-trend).

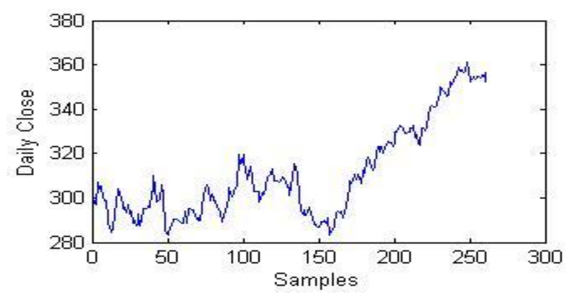
The selected stock price datasets are then checked for chaos using the test as described in [18] and [19]. The obtained values of λ are presented in Table 2. It was observed that all the stocks considered in the present study exhibit deterministic behavior and hence, successful prediction is possible.



(a)



(b)



(c)

Fig. 2. Trends in selected stock datasets (a) GSK (b) SBI (c) SBY

Table 2 Selected stocks and their Lambda (λ) values

Stock	Exchange	λ
	India	
NTPC	NSE	-8.43
ICICI		-9.08
SBI		-8.24
TCS		-9.08
	US	
Citibank	NASDAQ	-1.83
Google		-8.63
Microsoft		-0.40
Royal Dutch Shell	NYSE	-8.64

UK		
GSK	LSE	-8.86
RBS		-1.83
Rolls Royce Holding		-12.21
Sainsbury		-8.43

The optimal parameters obtained for regression based approach using GA are presented in Table 3

Table 3 Optimal parameters for regression based approach

Stock	Neurons	Alphabets
India		
NTPC	20	3
ICICI	100	3
SBI	50	3
TCS	100	3
US		
Citibank	50	3
Google	30	3
Microsoft	80	3
Royal Dutch Shell	50	3
UK		
GSK	10	3
RBS	30	3
Rolls Royce Holding	60	3
Sainsbury	70	3

The optimal parameters obtained for classification based approach are presented in Table 4

Table 4 Optimal parameters for classification based approach

Stock	Neurons	Alphabets
India		
NTPC	10	3
ICICI	60	3
SBI	40	3
TCS	40	3
US		
Citibank	10	3
Google	20	3
Microsoft	10	3
Royal Dutch Shell	60	3
UK		
GSK	10	3
RBS	10	3
Rolls Royce Holding	10	3
Sainsbury	20	3

The results obtained using both approaches are presented in Tables 5 and 6. The two parameters used for evaluating the effectiveness of the regression based forecasting system are Mean Absolute Percentage Error (MAPE) and the percentage accuracy in stock level prediction (ACC).

Table 5. Results-Regression

Stock	Training		Testing	
	MAPE	**ACC**	**MAPE**	**ACC**
India				
NTPC	7.12	86.32	0	100
ICICI	6.33	88.43	0	100
SBI	5.13	91.03	0	100
TCS	7.56	87.04	0	100
US				
Citibank	6.81	89.33	0	100
Google	4.89	92.00	0	100
Microsoft	4.22	90.67	12.67	76.00
Royal Dutch Shell	6.44	88.44	13.33	68.00
UK				
GSK	11.84	77.38	12.5	83.33
RBS	2.34	96.83	0	100
Rolls Royce Holding	3.71	91.82	0	100
Sainsbury	9.33	85.04	0	100

Table 6. Results-classification (% accuracy)

Stock	Training	Testing
India		
NTPC	86.32	100
ICICI	88.43	100

SBI	91.03	100
TCS	87.04	100
US		
Citibank	89.33	100
Google	92.00	100
Microsoft	90.67	76.00
Royal Dutch Shell	88.44	68.00
UK		
GSK	77.38	83.33
RBS	96.83	100
Rolls Royce Holding	91.82	100
Sainsbury	85.04	100

4. CONCLUSIONS

A novel GA-optimized SAX- ANN based Stock Level Prediction System was proposed in this study and its performance was empirically validated on twelve different stocks drawn from four different stock markets. It was observed that the proposed system is capable of learning patterns from the historical stock price data to forecast one-day-ahead stock price trends. This is evident from the out-of-sample results obtained (as presented in Tables 5 and 6). It is also observed that both the approaches presented in the study, result in similar levels of accuracy. This could be attributed to the fact that in both the approaches, the dataset can take only a finite set of values (members of the alphabet). The proposed system can also work on intraday (streaming) stock price data and may be used as a building block for an algorithmic trading system.

REFERENCES

[1] Fama, E. F. 1970. Efficient capital markets: A review of theory and empirical work, Journal of Finance, 25, 383–417.

[2] Fama, E.F. 1965. Random walks in stock market prices' Financial Analysts Journal, Vol. 21 No.5, 55-59.

[3] Atsalakis, G. S. and Valavanis K. P. 2009. Surveying stock market forecasting techniques – part II: soft computing methods, Expert Systems with Applications, Vol.36, 5932–5941.

[4] Nair, B.B. and Mohandas, V.P. 2015. Artificial intelligence applications in financial forecasting- A survey and some empirical results, Intelligent Decision Technologies. Accepted for publication.

[5] Nair, B.B. and Mohandas, V. P. 2015. An Intelligent Recommender System for Stock Trading, Intelligent Decision Technologies. Accepted for publication.

[6] Eng, W. F.1988. Technical Analysis of Stocks, Options and Futures: Advanced Trading Systems and Techniques, Irwin Professional Publishing.

[7] Nair,B.B., Mohandas,V.P., Sakthivel,N.R., Nagendran,S., Nareash,A., Nishanth,R., Ramkumar,S., Kumar,D.M. Application of hybrid adaptive filters for stock market prediction. In proceedings of the International Conference on Communication and Computational Intelligence (INCOCCI 2010), 443-447.

[8] Nair, B.B., Patturajan, M., Mohandas, V.P., Sreenivasan and R.R. Predicting the BSE Sensex: Performance comparison of adaptive linear element, feed forward and time delay neural networks, In proceedings of the International Conference on Power, Signals, Controls and Computation (EPSCICON 2012), 1-5.

[9] Nair, B.B., Minuvarthini, M., Sujithra, B., Mohandas, V.P. Stock market prediction using a hybrid neuro-fuzzy system. In proceedings of the International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom 2010), 243-247.

[10] Nair, B.B, Sai, S.G., Naveen, A.N., Lakshmi, A., Venkatesh, G.S. and Mohandas, V.P. 2011. A GA-artificial neural network hybrid system for financial time series forecasting, Information Technology and Mobile Communication, Springer Berlin Heidelberg, 499-506.

[11] Nair, B.B., Mohandas, V.P. and Sakthivel, N.R. 2010. A decision tree—rough set hybrid system for stock market trend prediction, International Journal of Computer Applications, Vol. 6, No.9, 1-6.

[12] Nair, B.B., Dharini, N.M. and Mohandas, V.P. 2010. A Stock Market Trend Prediction System Using a Hybrid Decision Tree-Neuro-Fuzzy System. In proceedings of the International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom-2010), 381-385.

[13] Nair, B.B., Mohandas, V.P and Sakthivel, N.R. 2010. A genetic algorithm optimized decision tree-SVM based stock market trend prediction system, International Journal on Computer Science and Engineering, Vol. 2, No. 9, 2981-2988.

[14] Nair, B.B., Mohandas, V.P., Sakthivel, N.R. 2011.Predicting stock market trends using hybrid ant–colony–based data mining algorithms: an empirical validation on the Bombay Stock Exchange, International Journal of Business Intelligence and Data Mining, Vol. 6, No. 4, 362-381.

[15] Nair, B.B., Mohandas, V.P., Varun, G., Chaitanya, I., Krishna, K.S., Karthik, S.M., Kumar, B.J.V.2013. A Temporal Association Rule Mining Based Decision Support System for Stock Trading, International Research Journal of Finance and Economics, Vol.5, No.117, 67-79.

[16] Nair, B.B, Krishna, S.K., Singhi, S., Prasath V.R.R., Arvind,M. L., Abhimanyu, A. and Mohandas V.P. Improving Returns from the Markowitz Model using GA-An Empirical Validation on the BSE. In Proceedings of the International Conference on Advances in Civil Engineering (AETACE 2013).

[17] Keogh, E. and Lin, J.2005. Clustering of time-series subsequences is meaningless: implications for previous and future research, Knowledge and information systems, 154-177.

[18] Saïda, A.B. and Litimi, H. 2013. High level chaos in the exchange and index markets, Chaos, Solitons & Fractals, Vol. 54, 90–95.

[19] Saïda, A.B. 2014. Noisy chaos in intraday financial data: Evidence from the American index, Applied Mathematics and Computation, Vol. 226, 258–265.

[20] Percival, D.B. and Walden, A.T.2000. Wavelet Methods for Time Series Analysis, Cambridge University Press.