Journal of International Technology and Information Management

Volume 20
Issue 1 *Double Issue 1/2*

Article 3

2011

A Hybrid Machine Learning System for Stock Market Forecasting

Lokesh Kumar
Babu Banarasi Das National Institute of Technology and Management

Anvita Pandey
Babu Banarasi Das National Institute of Technology and Management

Saakshi Srivastava
Babu Banarasi Das National Institute of Technology and Management

Manuj Darbari
Babu Banarasi Das National Institute of Technology and Management

Follow this and additional works at: https://scholarworks.lib.csusb.edu/jitim

Part of the Management Information Systems Commons

Recommended Citation

Kumar, Lokesh; Pandey, Anvita; Srivastava, Saakshi; and Darbari, Manuj (2011) "A Hybrid Machine Learning System for Stock Market Forecasting," *Journal of International Technology and Information Management*: Vol. 20: Iss. 1, Article 3.
DOI: https://doi.org/10.58729/1941-6679.1099
Available at: https://scholarworks.lib.csusb.edu/jitim/vol20/iss1/3

This Article is brought to you for free and open access by CSUSB ScholarWorks. It has been accepted for inclusion in Journal of International Technology and Information Management by an authorized editor of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.

A Hybrid Machine Learning System for Stock Market Forecasting

Lokesh Kumar
Anvita Pandey
Saakshi Srivastava
Manuj Darbari
Babu Banarasi Das National Institute of Technology and Management
INDIA

ABSTRACT

A hybrid machine learning system based on Genetic Algorithm (GA) and Time Series Analysis is proposed. In stock market, a technical trading rule is a popular tool for analysts and users to do their research and decide to buy or sell their shares. The key issue for the success of a trading rule is the selection of values for all parameters and their combinations. However, the range of parameters can vary in a large domain, so it is difficult for users to find the best parameter combination. In this paper, we present the Genetic Algorithm (GA) to overcome the problem in two steps. First, setting a sub-domain of the parameters with GA. Second, finding a near optimal value in the sub domain with GA and Time Series Analysis in a very reasonable time.

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 profit (Chen, Leung, & Daouk, 2003).

TIME SERIES ANALYSIS

Forecasting, or predicting, is an essential tool in any decision-making process. Its uses vary from determining inventory requirements for a local shoe store to estimating the annual sales of video games. The quality of the forecasts management can make is strongly related to the information that can be extracted and used from past data. Time-series analysis is one quantitative method we used to determine patterns in data collected over time. Time-series analysis is used to detect patterns of change in statistical information over regular intervals of time. We project these patterns to arrive at an estimate for the future. Thus Time-series analysis helps us cope with uncertainty about the future.

GENETIC ALGORITHM

Genetic algorithms are search algorithms based on the mechanics of natural selection and natural genetics. They combine survival of the fittest among string structures with a structured yet randomized information exchange to form a search algorithm with some of the innovative flair of human search. In every generation, a new set of artificial creatures (strings) is created using bits and pieces of the fittest of the old; an occasional new part is tried for good measure. While randomized, genetic algorithms are no simple random walk. They efficiently exploit historical information to speculate on new search points with expected improved performance.

Generally, genetic operations include: " crossover" , "mutation" and " selection" .

"Crossover" operator: Suppose $S1=\{s11, s12, \ldots , s1n\}$, $S2=\{s21, s22, \ldots , s2n\}$, are
two chromosomes, select a random integer number $0\ r\ n$, $S3$, $S4$ are offspring of crossover($S1$, $S2$),
$S3=\{si \mid \text{if } i\ r,\ si\ S1, \text{ else } si\ S2\}$,
$S4=\{si \mid \text{if } i\ r,\ si\ S2, \text{ else } si\ S1\}$.

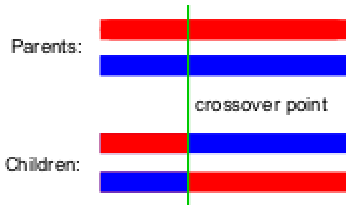
"Mutation" operator: Suppose a chromosome $S1=\{s11, s12, \ldots , s1n\}$, select a
random integer number $0\ r\ n$, $S3$ is a mutation of $S1$,
$S3=\{si \mid \text{if } i\ r, \text{ then } si=s1i, \text{ else } si = \text{random}(s1i)\}$.

"Selection" operator: Suppose there are *m* individuals,
If m is even: ***no change***
If m is odd: i. take mean of the dataset.
ii. Now, mutate each individual value with this mean
iii. Match each newly generated population to the mean value if the bit value (up to r- random value used for mutation), then reject this data set. i.e. Consider other one.

In the proposed system, we use the single point crossover and flip bit(bit string) mutation.

Single Point Crossover: A single crossover point on both parents' organism strings is selected. All data beyond that point in either organism string is swapped between the two parent organisms. The resulting organisms are the children:

Figure 1: Single point crossover.



Flip-Bit Mutation: The mutation of bit strings ensue through bit flips at random positions.

Example:

1 0 1 0 0 1 0

↓

11 0 1 0 1 1 0

The probability of a mutation of bits is $\frac{1}{l}$, where *l* is the length of the binary vector. Thus, a mutation rate of 1 per mutation and individual selected for mutation is reached.

RELATED RESEARCH

We are interested in the profitability of trading rules rather than accuracy of prediction itself - the economics literature is concerned with market efficiency instead of prediction accuracy, and the ability of a trading strategy to make consistent excess profits is a strong sing of market inefficiency. Therefore, although are algorithms will give us ‘up’ or ‘down’ predictions, we will map these predictions onto a trading strategy. Specifically, if our algorithms predict up, we buy or hold the stock (depending on if we already own it) If our algorithms predict down, we sell the stock or hold a cash position.

LITERATURE SURVEY

A number of artificial intelligence and machine learning techniques have been used over the past decade to predict the stock market. However, stock market prediction networks have also been implemented using genetic algorithms, recurrent networks, and modular networks. This section discusses some of the network architectures used and their effect on performance. Back propagation networks (Tan, Prokhorov, & Wunsch, 1995) are the most commonly used network because they offer good generalization abilities and are relatively straightforward to implement. Although it may be difficult to determine the optimal network configuration and network parameters, these networks offer very good performance when trained appropriately.

The JSE-system was a back propagation network designed using a genetic algorithm. The genetic algorithm allowed the automated design of the neural network, and determined that the optimal network configuration was one hidden layer with 21 nodes. Genetic algorithms are especially useful where the input dimensionality is large. They allowed the network developers to automate network configuration without relying on heuristics or trial-and-error. The Tokyo stock prediction system (Kimoto, Asakawa, Yoda, & Takeoka, 1990) was a modular neural network consisting of 4 back propagation networks trained on different data items. Many other stock market prediction systems are also based on the back propagation network (Refenes, Zapranis, & Francis, 1995).

Recurrent network architectures are the second most commonly implemented architecture. The motivation behind using recurrence is that pricing patterns may repeat in time. A network which remembers previous inputs or feedbacks previous outputs may have greater success in determining these time dependent patterns. There are a variety of such networks which may have recurrent connections between layers, or remember previous outputs and use them as new inputs to the system (increases input space dimensionality).

The performance of these networks are quite good. A recurrent network model was used in (Saad, Prokhorov, & Wunsch, 1996). A self-organizing system was also developed by Wilson (Kamijo & Tanigawa, 1993) to predict stock prices. The self-organizing network was designed to construct a nonlinear chaotic model of stock prices from volume and price data. Features in the data were automatically extracted and classified by the system. The benefit in using a self-organizing neural network is it reduces the number of features (hidden nodes) required for pattern classification, and the network organization is developed automatically during training.

Wilson used two self-organizing neural networks in tandem; one selected and detected features of the data, while the other performed pattern classification. Over fitting and difficulties in training were still problems in this organization.

An interesting hybrid network architecture was developed in (Kamijo & Tanigawa, 1993), which combined a neural network with an expert system. The neural network was used to predict future stock prices and generate trading signals. The expert system used its management rules and formulated trading techniques to validate the neural network output. If the output violated known principles, the expert system could veto the neural network output, but would not generate an output of its own. This architecture has potential because it combines the nonlinear prediction of neural networks with the rule-based knowledge of expert systems. Thus, *the combination of the two systems offers superior knowledge and performance*.

There is no one correct network organization. Each network architecture has its own benefits and drawbacks. Back propagation networks are common because they offer good performance, but are often difficult to train and configure. Recurrent networks offer some benefits over back propagation networks because their "memory feature" can be used to extract time dependencies in the data, and thus enhance prediction. More complicated models may be useful to reduce error or network configuration problems, but are often more complex to train and analyze.

THE STOCK PREDCTION PROBLEM

The stock market direction problem is modeled as a two class classification problem. The directions are categorized as 0 & 1 in the data (Chen, 2002). 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 (Goldberg, 2005; Neely, Weller, & Dittmar, 1997). In the Indian stock market, we have chosen the stocks of; Tata consultancy services (TCS).

PROPOSED SYSTEM

Input dataset

As we have chosen the stock of TCS and Infosys for our prediction, we collected 259 days trading values, which includes (the opening, highest, lowest and closing values of the stock price for each day). As we aware of the fact that: the opening price is based on the closing price of previous day. So, to make a healthy prediction for a opening price we are considering the closing price and for closing price we are considering the open price of the stock. While the lowest and the highest values remains same.

Genetic algorithm:

Here, we are taking the closing price values of previous week to make the prediction of open price for the first day of the next week.
Days(X): 07,08,09,10,11,14,15,16,17

ClosingPrice(Y):

1132.05,1128.45,1098,1090.65,1089.4,1113.55,1096.8,1098.3,1109.2,1090.4
Round off: 1132,1128,1098,1091,1089,1114,1097,1098,1109,1090
Here, the value of m is even: so we can simply apply GA i.e. performing single point crossover and then applying flip bit mutation using a random function.

After Performing GA:

New Values: 1004,1128,1074,1091,833,994,1001,994,1104,1066
Applying Time-Series Analysis

Time-Series Analysis:

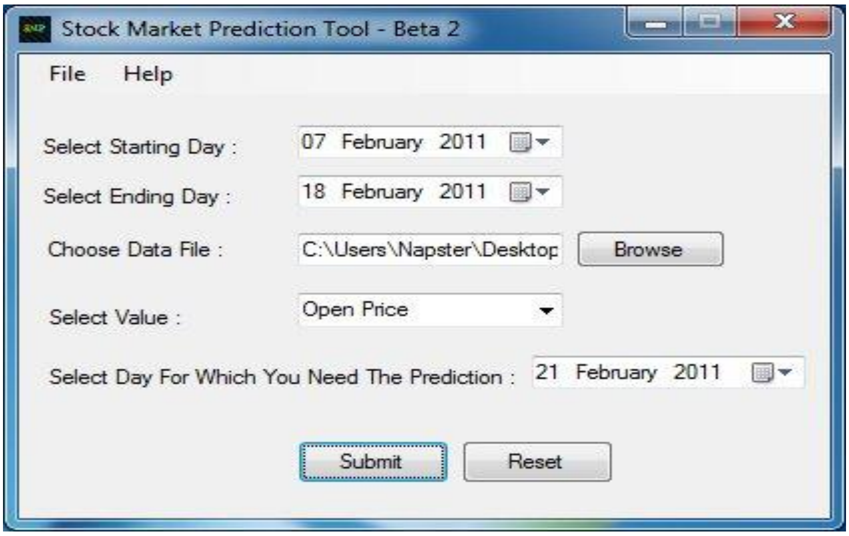
We get: +0.004305x (Neglecting –ve sign)*

*Neglecting "–ve" sign, while predicting the opening price from closing price because open price is always more than the closing price of previous day. But, when predicting closing price from open price we change the sign from "+ve" to "–ve" because closing price is always lower than the open price of next day. While predicting lowest and highest price, the sign does not change.

FINAL RESULT

From the above steps, we are able to derive the following result:
For Day 21: x=(21-12.5) where, 12.5 is the avg. value of days
x=8.5

Figure 2: Stock market prediction tool (snapshot_1).



Select Starting Day : 07 February 2011

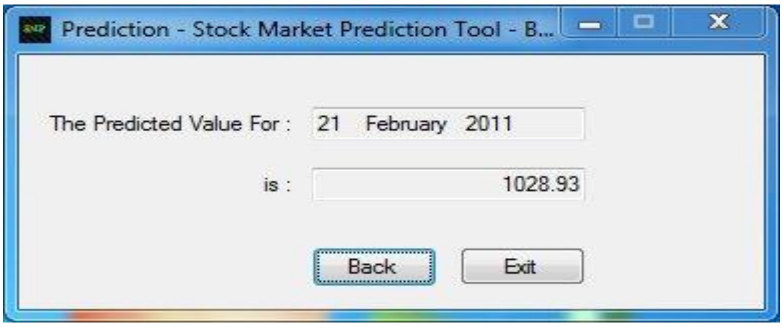
Select Ending Day : 18 February 2011

Choose Data File : C:\Users\Napster\Desktop

Select Value : Open Price

Select Day For Which You Need The Prediction : 21 February 2011

Figure 3: Stock market prediction tool (snapshot_2).



The Predicted Value For : 21 February 2011

is : 1028.93

So, the predicted value for day 21 is: 1028.93. Whereas, the actual value for day 21 is: 1090. As we know predicting the direction of the stock price will generate higher profit. So, if we focus on the direction of the prediction.

In this case, the previous day value is 1108.75 and the present day value is 1090, whereas we predicted the value of present day is 1028.93. Now, while focusing on the direction of the stock price movement we can see that our system predicted a down fall in the stock price and in real the stock price falls down.

HYPOTHESIS TESTING

Data are an essential input to the effective designing and implementation of transport systems by allowing the calibration of model which yields insights into the process at work in the system or to predict how the system is likely to perform. Design of the system is intimately associated with the definition of the experimental hypothesis.

There is significant variation in the type of industries/organizations. So, to check whether the prediction is significant or only due to chance we use chi-square test (Levin & Rubin) by consider the following hypothesis:

H_0 = our model is correct.
H_1 = our model is incorrect.

Calculation of χ^2

F_0	F_e	(f_0-f_e)	$(f_0-f_e)^2$	$(f_0-f_e)^2/f_e$
1106.25	1108.75	-2.5	6.25	0.00563
1087	1090.4	-3.4	11.56	0.01060
68.4	67.74	0.66	0.4356	0.00643
65.23	66.32	-1.09	2.18	0.03287

$$\chi^2 = \sum [(f_0 - f_e)^2/f_e]$$

Substituting the values in equation, we get:

$$\chi^2 = \mathbf{0.0553}$$

We compared the observed values of χ^2 with critical value of χ^2 and follow the rules of hypothesis:

$\chi^2_{\text{observed}} < \chi^2_{\text{Critical}}$ => Accept the Null Hypothesis
and if,

$\chi^2_{\text{observed}} > \chi^2_{\text{Critical}}$ => Reject the Null Hypothesis

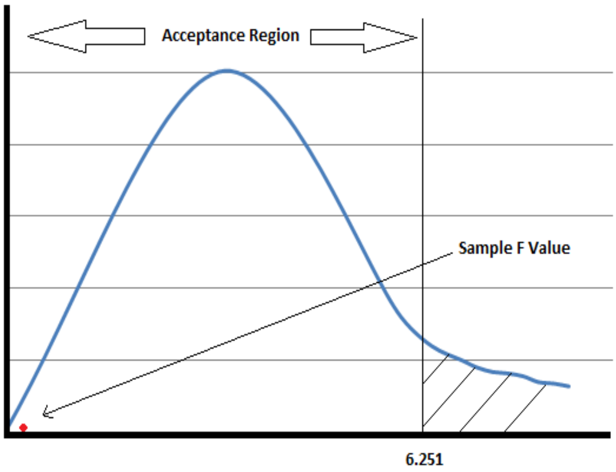
Now, calculating the Degree of Freedom, we get:

dF = 3

Tabulated value of χ^2 at 10% level of significance is **6.251**.

Since the calculated value of $\chi^2 = 0.0553 < 6.251$ the tabulated value, the null hypothesis H_0 is accepted at 10% level of significance.

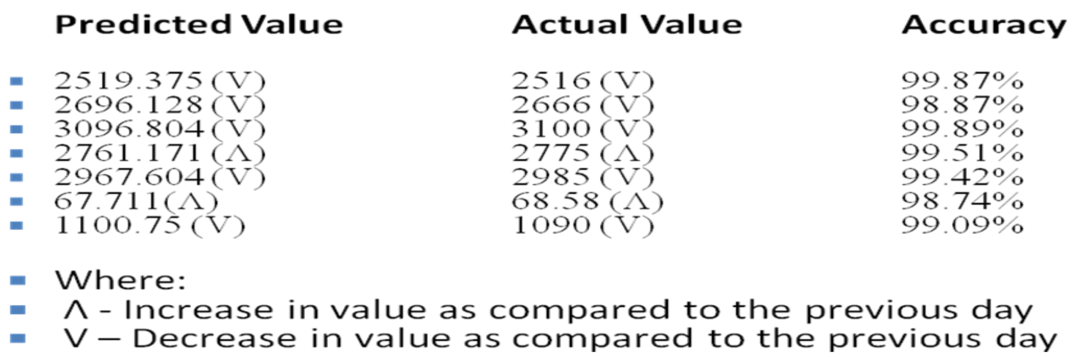
Figure 4: Sample acceptance region.



CONCLUSION

A hybrid model is proposed called the GA-TSA system for predicting the future direction of stock prices. The results showed that the GA and Time Series helped in improving the performance of the Prediction 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.

Effectiveness of SMP Tool

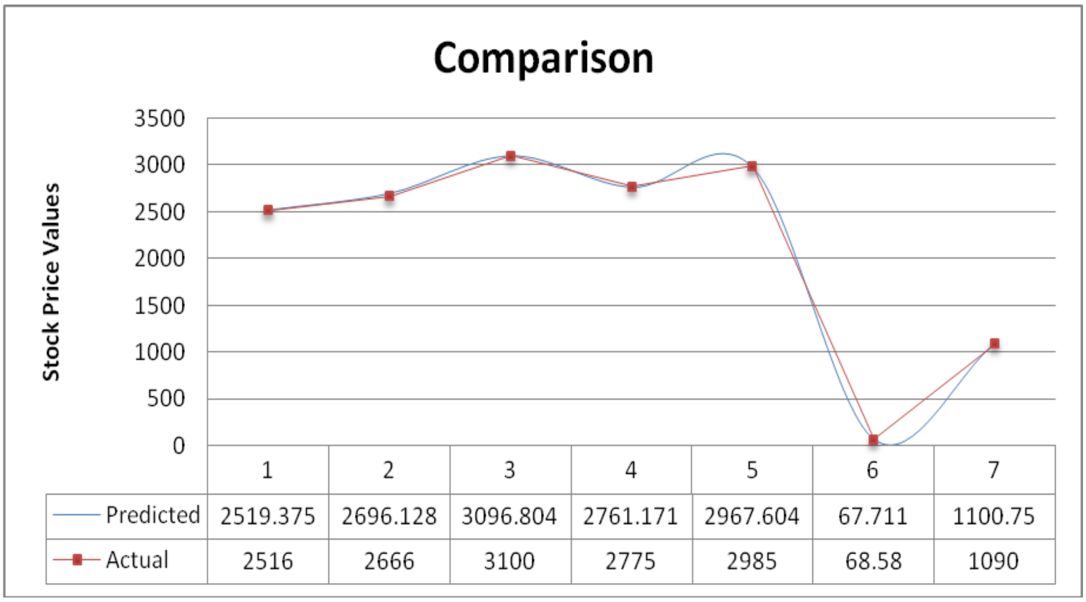


Predicted Value	Actual Value	Accuracy
2519.375 (V)	2516 (V)	99.87%
2696.128 (V)	2666 (V)	98.87%
3096.804 (V)	3100 (V)	99.89%
2761.171 (Λ)	2775 (Λ)	99.51%
2967.604 (V)	2985 (V)	99.42%
67.711(Λ)	68.58 (Λ)	98.74%
1100.75 (V)	1090 (V)	99.09%

- Where:
- Λ - Increase in value as compared to the previous day
- V – Decrease in value as compared to the previous day

Result comparison for effectiveness

Figure 4: Comparison for effectiveness.



	1	2	3	4	5	6	7
Predicted	2519.375	2696.128	3096.804	2761.171	2967.604	67.711	1100.75
Actual	2516	2666	3100	2775	2985	68.58	1090

REFERENCES

Chen, A. S., Leung, M. T., & Daouk, H. (2003). Application of neural networks to an emerging financial market: forecasting and trading the Taiwan stock index. *Computers and Operations Research*, 30, 901-923.

Chen, S. H. (2002). *Genetic Algorithms and Genetic Programming in Computational Finance.* Boston, MA: Kluwer.

Darbari, M., Srivastava, A. K., & Medhavi, S. (2007). Application of UML modelling urban traffic system using producer consumer theory to generate process algebra model. *Journal of International Technology and Information management,* 16(4), 75-84.

Darbari, M. & Karn, B. (2008). Enterprise modeling using unified framework supporting distributed object computing. *Journal of International Technology and Information Management,* 17(3/4), 205-218.

Darbari, M., Srivastava, A. K., & Medhavi, S. (2009). Formal verification of urban traffic system using the concept of fuzzy workflow simulation. *Journal of International Technology and Information Management,* 18(1), 59-74.

Darbari, M., Srivastava, A. K., & Medhavi, S. (2009). PENTRAL: pattern based logic language. *Journal of International Technology and Information management,* 18(3/4), 385-394.

Goldberg, D. E. (2005). *Genetic Algorithms in Search, Optimization, and Machine Learning*, The University of Alabama, Ninth Edition.

Kamijo, K., & Tanigawa, T. (1993). Stock price pattern recognition: A recurrent neural network approach, *NeuralNetworks in Finance and Investing*, 357–370.

Kim, K. J. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications*, 19(2), 125–132.

Kimoto, T., Asakawa, K., Yoda, M., & Takeoka, M. (1990). Stock market prediction system with modular neural networks. In *Proceedings of the International Joint Conference on Neural Networks*, 1, 1–6.

Lawrence, R. (1997). Using Neural Networks to Forecast Stock Market Prices. https://people.ok.ubc.ca/rlawrenc/research/Papers/nn.pdf

Levin, R. I. & Rubin, D. S. (1998). *Statistics for Management*, seventh edition.

Neely, C., Weller, P. & Dittmar, R. (1997). Is technical analysis in the foreign exchange market profitable? A geneticprogramming approach. *Journal of Financial and Quantitative Analysis*. 32, 405-26.

Refenes, A. P., Zapranis, A. D., & Francis, G. (1995). Modelling stock returns in the framework of APT: A comparative study with regression models. *Neural Networks in the Capital Markets*, 101–137.

Tan, H., Prokhorov, D., & Wunsch, D. (1995). Probabilistic and time-delay neural-network techniques for conservative short-term stock trend prediction. *Proceeding of the World Congress of Neural Networks*, Washington, D.C.

Saad, E., Prokhorov, D., & Wunsch, D. (1996). Advanced neural-network training methods for low false alarm stock trend prediction, in *Proceedings of the IEEE International Conference of Neural Networks,* Washington, D.C., June.