DETECTION OF STATISTICAL ARBITRAGE USING MACHINE LEARNING TECHNIQUES IN INDIAN STOCK MARKETS

A.U.S.S PRADEEP (DEEPU@IITK.AC.IN), SOREN GOYAL (SOREN@IITK.AC.IN)
ADVISOR: DR. AMITABHA MUKERJEE
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
IIT KANPUR ,INDIA
APRIL 15,2013

1. OBJECTIVE

The aim of the project is to analyze Arbitrage opportunities arising in the Indian stock markets modeled on the set of previous historical data using the following two techniques – Regression and Time Delay Neural Networks

2. INTRODUCTION

Before we describe the problem precisely, some background discussion about statistical arbitrage is necessary. "Statistical arbitrage refers to attempting to profit from pricing inefficiencies identified through mathematical models" (Patra & Fu, 2009). The basic assumption is that prices will move towards a historical average.

Consider a simple example of Arbitrage. If doll sells for Rs100 in Kanpur and for Rs200 in New Delhi, one could buy the doll in Kanpur and sell it in New Delhi for a profit of Rs100. Such a situation occurs in the market because information about prices takes time to travel from one place to another. But such a situation never lasts long, over time the prices in different places converge to a single value. In this case the price of the doll in Kanpur and New Delhi will converge to Rs150.

The price of a stock is decided by the market conditions, however at times it may happen that the stock may be mispriced. The reasons for this could lie in human error or market inefficiencies. In any case the price eventually tends to the correct market price. In this analysis we attempted to predict the market price of the target stock, using the prices of the stocks related to it. Then we compare this price with the listed price (Actual price) of the target stock to say if an arbitrage has occurred or not.

3. WORK DONE PREVIOUSLY:

Over more than half a century, much empirical research was done on testing the market efficiency, which can be traced to 1930's by Alfred Cowles, Many studies have found that stock prices are at least partially predictable. The method to test the existence of statistical arbitrage was finally described in the paper "Statistical arbitrage and tests of market efficiency" [4] by, S.Horgan, R.Jarrow, and M. Warachka published in 2002.And an improvement on the paper "An Improved test for Statistical arbitrage"[5] was published in 2011 by the same team which forms the basis for this project.

And further a paper "Machine Learning in Statistical Arbitrage" [1] (2009) tries to implement an approach of Support Vector Regression (SVR) and Principal component Analysis (PCA) to device a trading strategy to utilize the arbitrage opportunity the iShares FTSE/MACQ traded in the London Stock Exchange Market.

Also a paper "Statistical Arbitrage Stock Trading using Time Delay Neural Networks"(2004)[6] attempts to solve the problem whether TDNN architecture trained on the past history of stock data (NASDAQ dataset from 1975-2000)can accurately predict when to buy a stock.

4. MOTIVATION:

Arbitrage has the effect of causing prices in different markets to converge. [3] "The speed at which the convergence process occurs usually gives us a measure of the market efficiency".

Hence a thorough analysis of statistical arbitrage opportunities using the advanced learning techniques is essential in mapping the efficiency of current day Indian market.

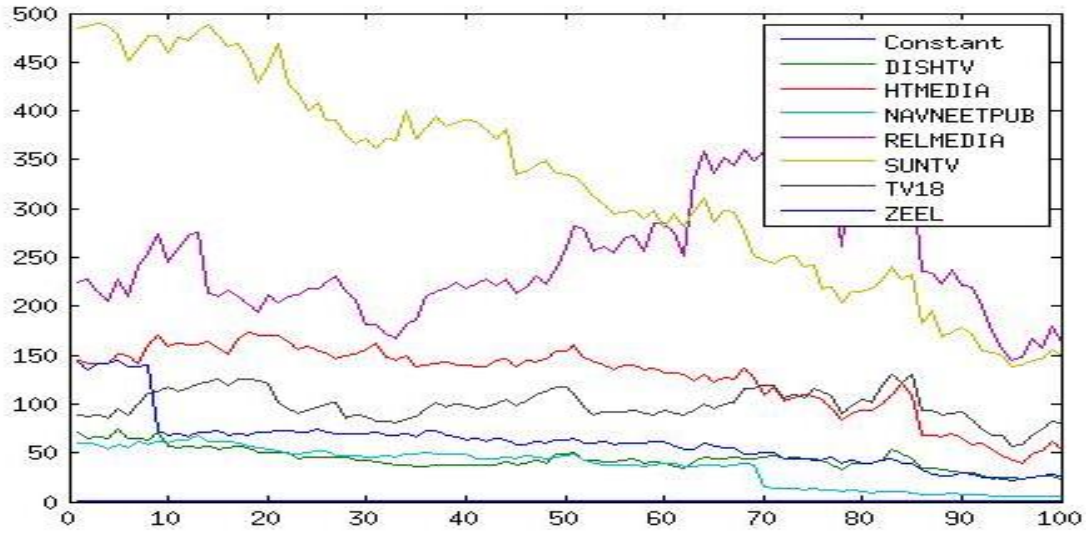
5. Detection of Arbitrage using Least Square Regression (Patra & Fu, 2009)

Target Stock: JAGRAN

COMPONENT STOCKS:

1. DISHTV
2. HTMEDIA
3. NAVNEETPUB
4. RELMEDIA
5. SUNTV
6. TV18
7. ZEEL

Prices Vs time for all chosen stocks



Procedure

Choosing the Stocks for analysis:

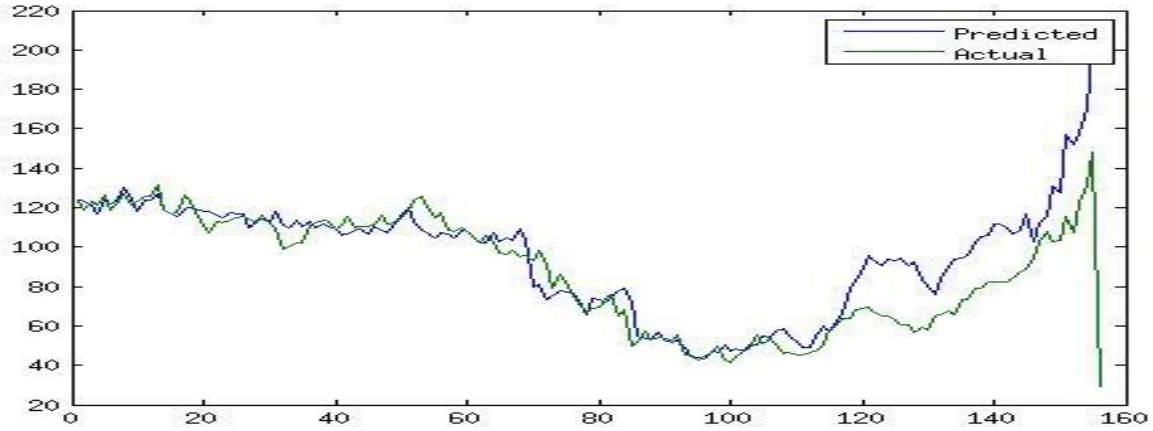
We chose the media sector for analysis, the decision was arbitrary. The 7 stocks chosen were the members of the NSE CNX MEDIA index. These stock will be later used to the model an index, which will mimic the variations of the member stocks. These stocks were chosen in particular because they best represented the conditions of the media sector.

The target stock was chosen as Jagran Media, as it was one of the lesser components of the CNX Media index and we were hopeful that it would show some dependence on the prices of the other stocks.

INITIAL ANALYSIS

We started off by first confirming that the, indeed our target stock was dependent on the index stocks. This was done in the following steps –

1. Data was collected from the historical database of NSE - from the year 2007 to 2010.
2. The data from 2007 to 2009 was used to generate a set of 7 coefficients using multi linear regression. The 7 index stocks linearly combined to give the target stock.
3. Using the coefficients and the data from beyond 2009 to 2010, we generate the index. This is then compared with the actual value of the Jagran stock.

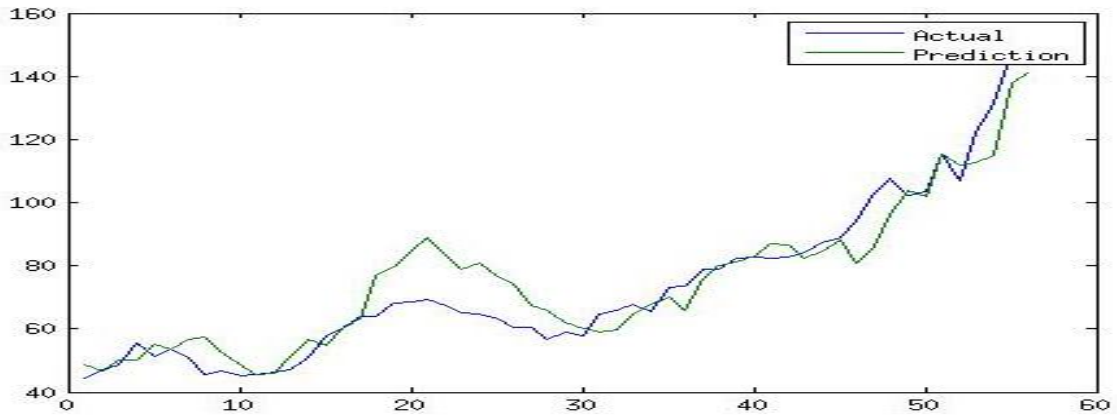


1 X-AXIS TIME IN WEEK, Y-AXIS PRICE OF THE STOCKS

It is seen that indeed the index is mimicking the variations of the actual stock. The sharp drop in the actual value in the end was due an uncorrected stock split.

MAIN ANALYSIS

NOW STARTING AGAIN FROM 2007 TO 2009 WE DYNAMICALLY FITTED THE DATA POINTS TO OBTAIN THE MARKET PRICE FOR THE NEXT DAY UNTIL THE YEAR 2010.



2 X- AXIS TIME IN WEEKS, Y-AXIS PRICE OF STOCK

6. Prediction Using Neural Networks:

To refine our approach and attain a better prediction we tried the time series model, historical data is collected and analyzed to produce a model that could understand the relations between the observed variables. The model is then used to predict future price value of the stock based on this time series. Artificial Neural networks can be used for statistical modeling and is an alternative to linear regression models, which are the most common approach for creating a predictive models.

"Neural networks have several advantages including less need for formal statistical training, ability to detect, implicitly, complex nonlinear relationships between dependent and independent variables, ability to detect any possible interactions between predictor variables and the existence of a wide variety of training algorithms"[1].

"Disadvantages of neural network include the nature of "black box" computing, inclination for memorize the data (network loses the ability to generalize), and the empirical nature of the model developed" [1]

A Neural network is always trained, so that a signal input can achieve a desired target. The system is them adjusted based on the comparisons made with the response and the desired target, until the network output matches the target. The below schematic shows how this supervised training in a network works.

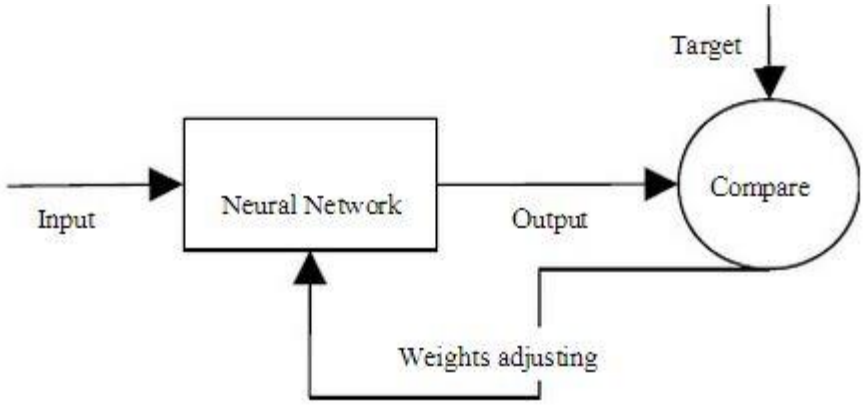


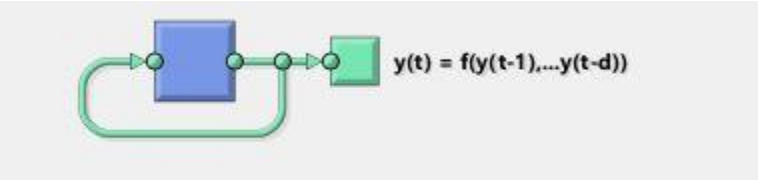
Fig. 2.1. Supervised training of a neural network.

Image source: "On a Model for predicting the exchange rate Euro-Leu with A NAR neural network", Dumitru Ciobanu.

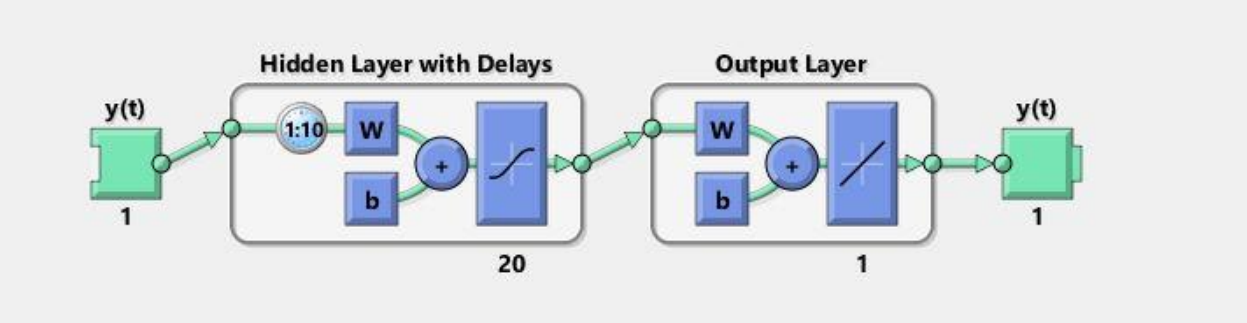
Neural networks are very good universal approximators and work best when used for modeling systems that have high tolerance to errors. "The strength of neural networks is their ability to accurately predict outcomes for complex problems. In tests of accuracy, compared with other approaches, neural networks are always able to get very good scores" (Berson, Smith & Thearling, 1999).

PROCEDURE:

In this method we tried to predict the future price of a given stock namely “ITC Ltd. - BSE” depending on the historical stock prices of the same stock.



We used a NAR (Non-linear Auto Regressive) Neural Network that uses 10 (d=10) past values of the stock price and predict the next value in the time series. The default structure of the neural network used consists of 20 hidden neurons with sigmoidal activation function and an output neuron with linear activation function.



TRAINING ALGORITHM:

Levenberg-Marquardt backpropagation was used, in this process errors are propagated backwards from the output layer toward the input while training. This is necessary because hidden units have no training target value that can be used, so they must be trained based on the errors from previous layers. The only layer that has a target value for comparison purpose is the output layer. As the errors are backpropagated through the nodes, the connection weights are changed. Training occurs until the errors in weights are sufficiently small to be accepted.

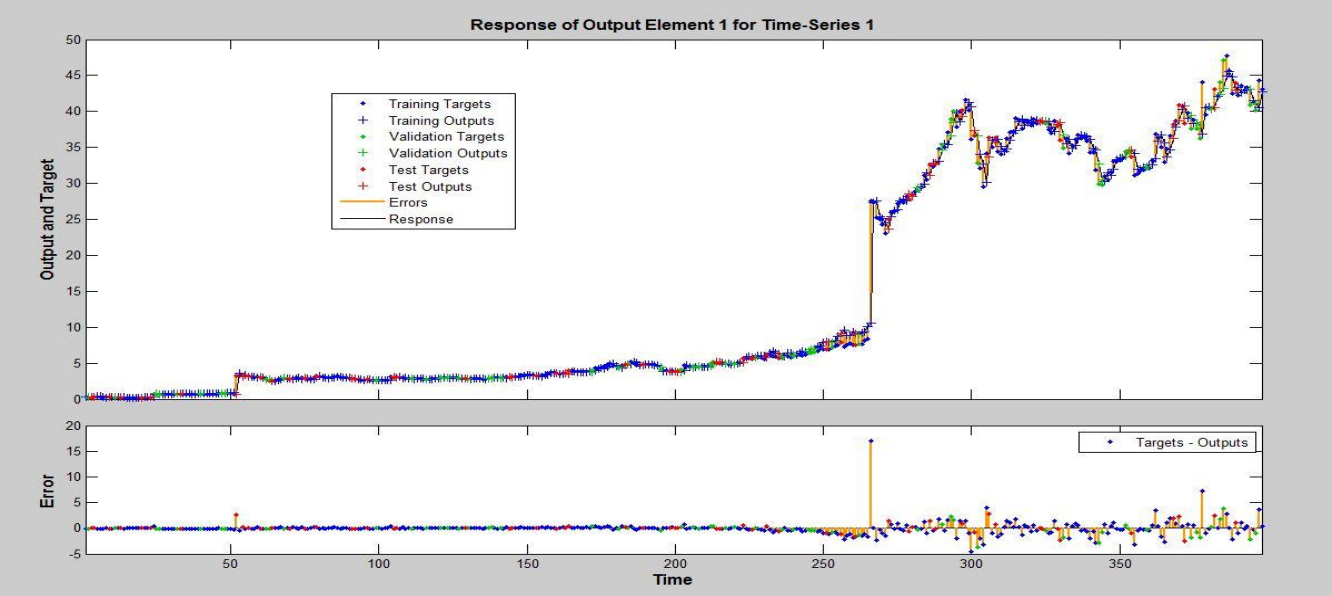
And lastly the data is divided in the following quantities 70% - Training, 15% - Validation and 15% - Testing .The performance is then estimated using the MSE-Mean Squared Error function.

Using the data we trained a data of stock prices at the end of the week, for 400 consecutive weeks (about 8 years from 2000-2008).And on this trained network we tried to predict the prices for the next 250 weeks and compared the accuracy by varying the size of the network.

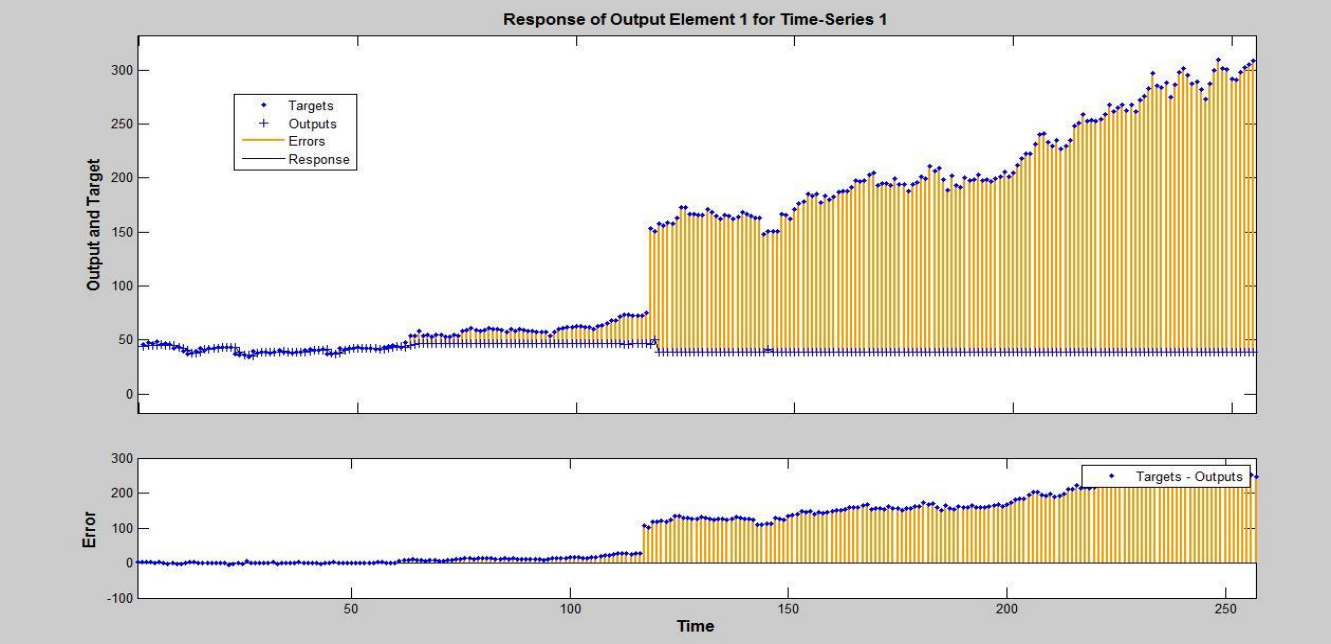
RESULTS:

1. For a network with 10 hidden neurons and delay of 2 :

Training data:

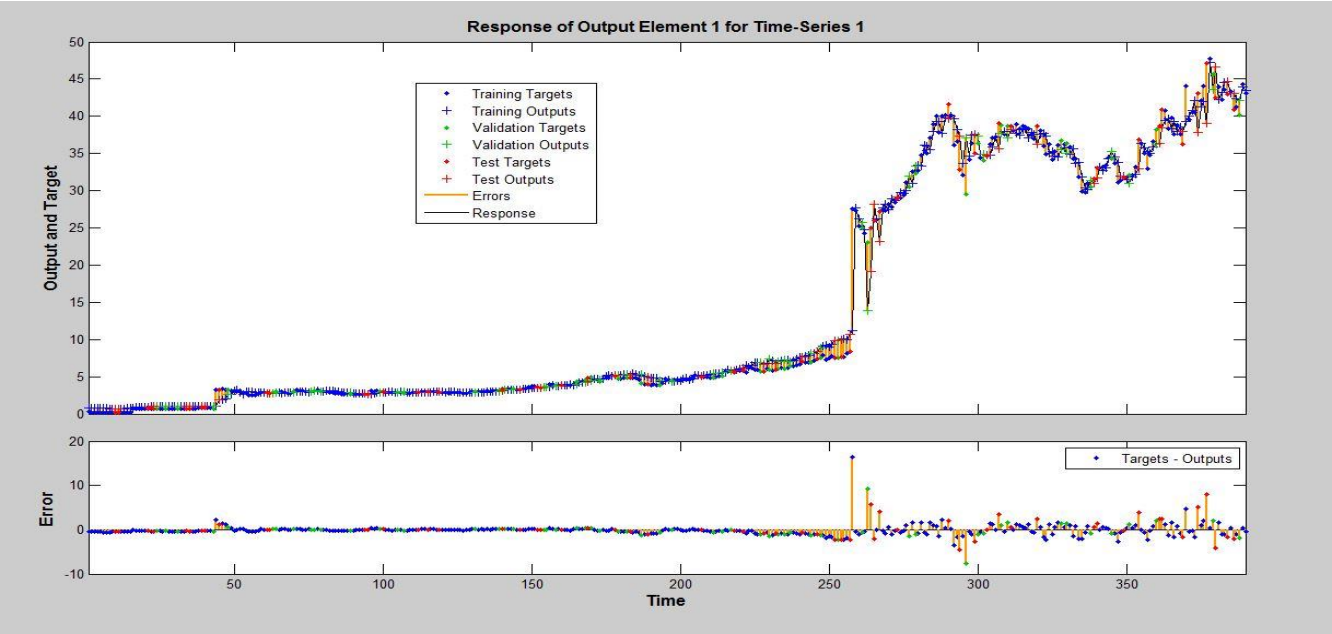


Future Prediction:

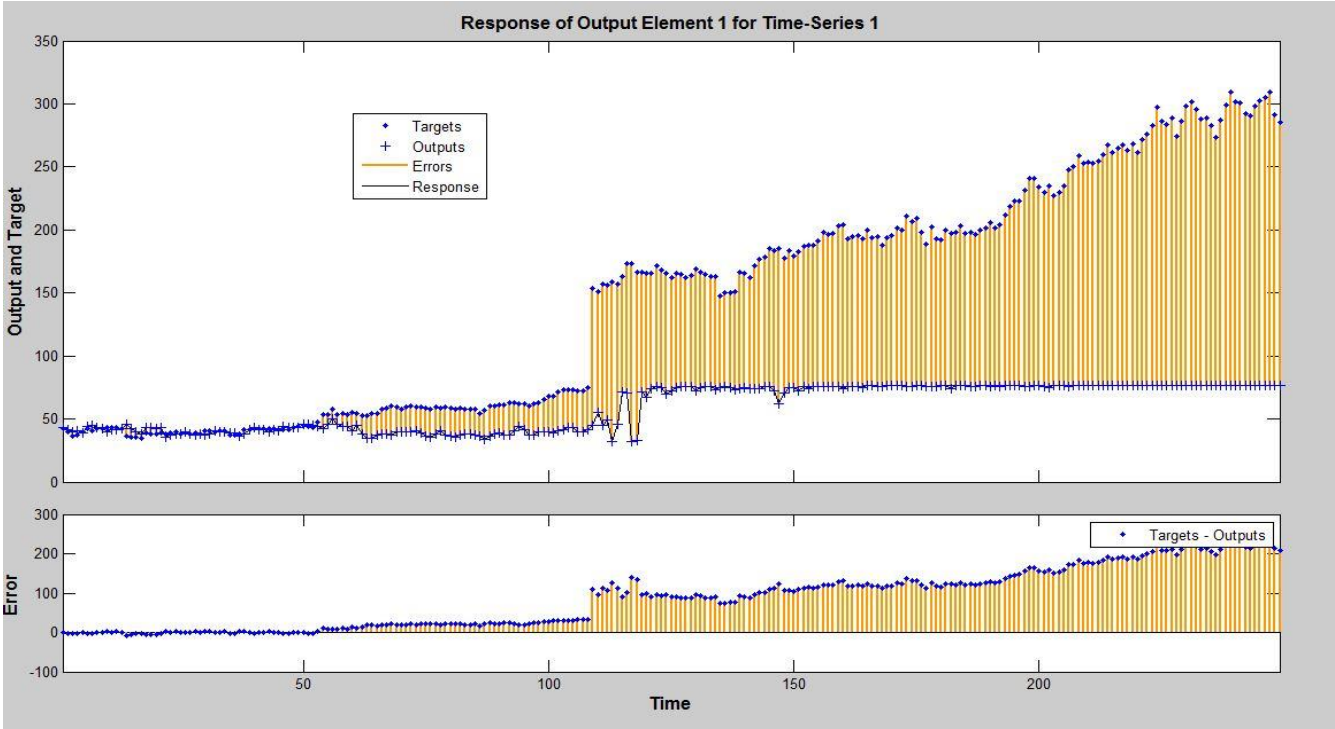


2. For a network with 10 hidden neurons and delay of 10:

Training data:

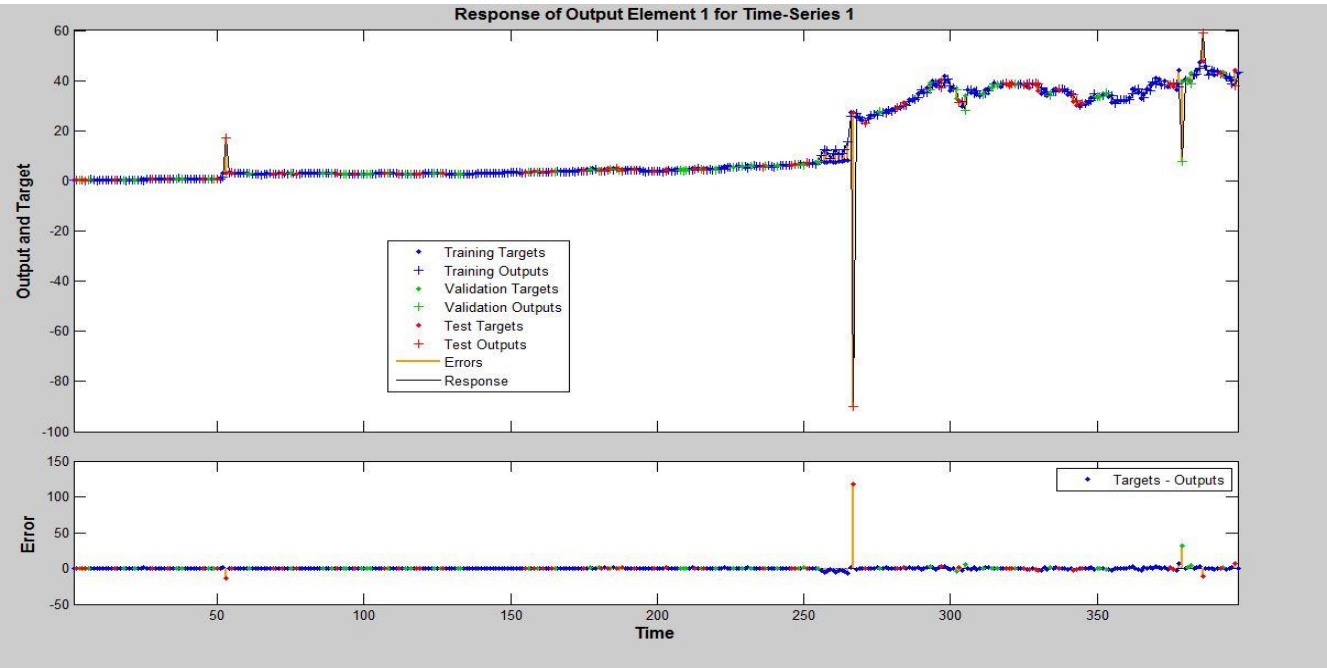


Future Prediction:

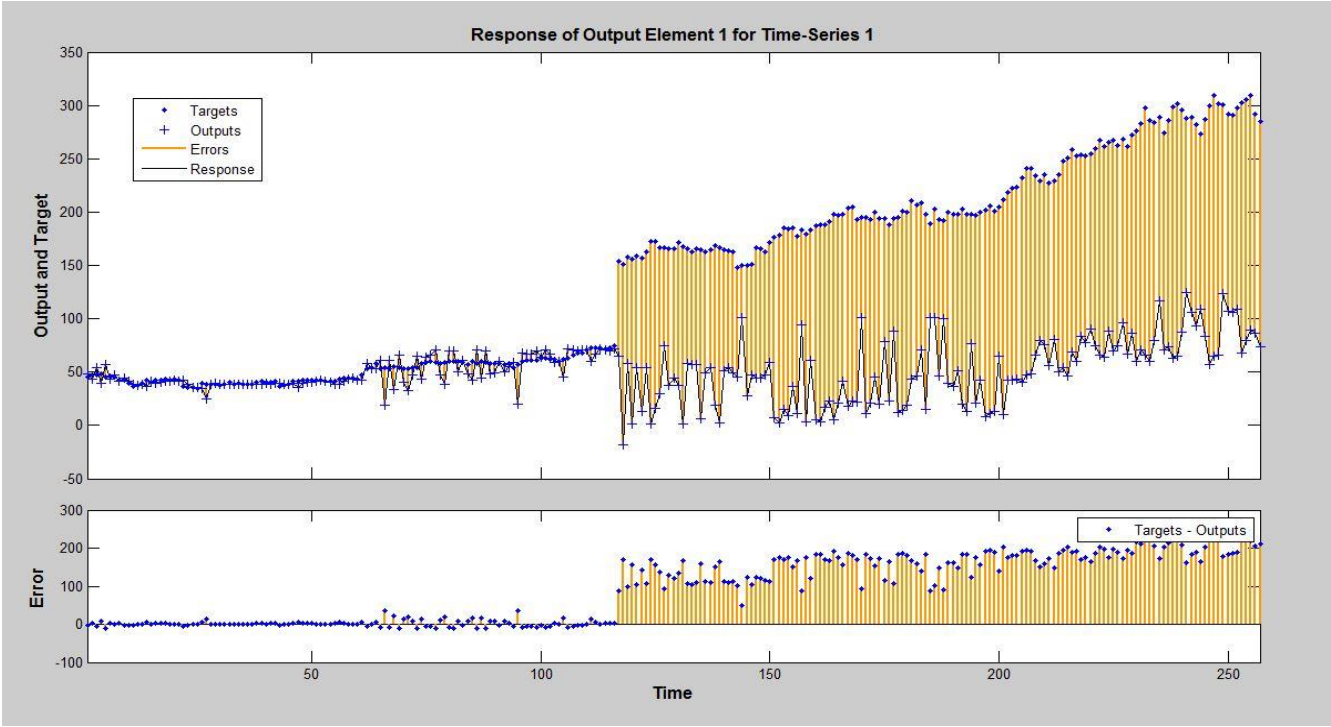


3. For a network with 100 hidden neurons and delay of 2:

Training data:

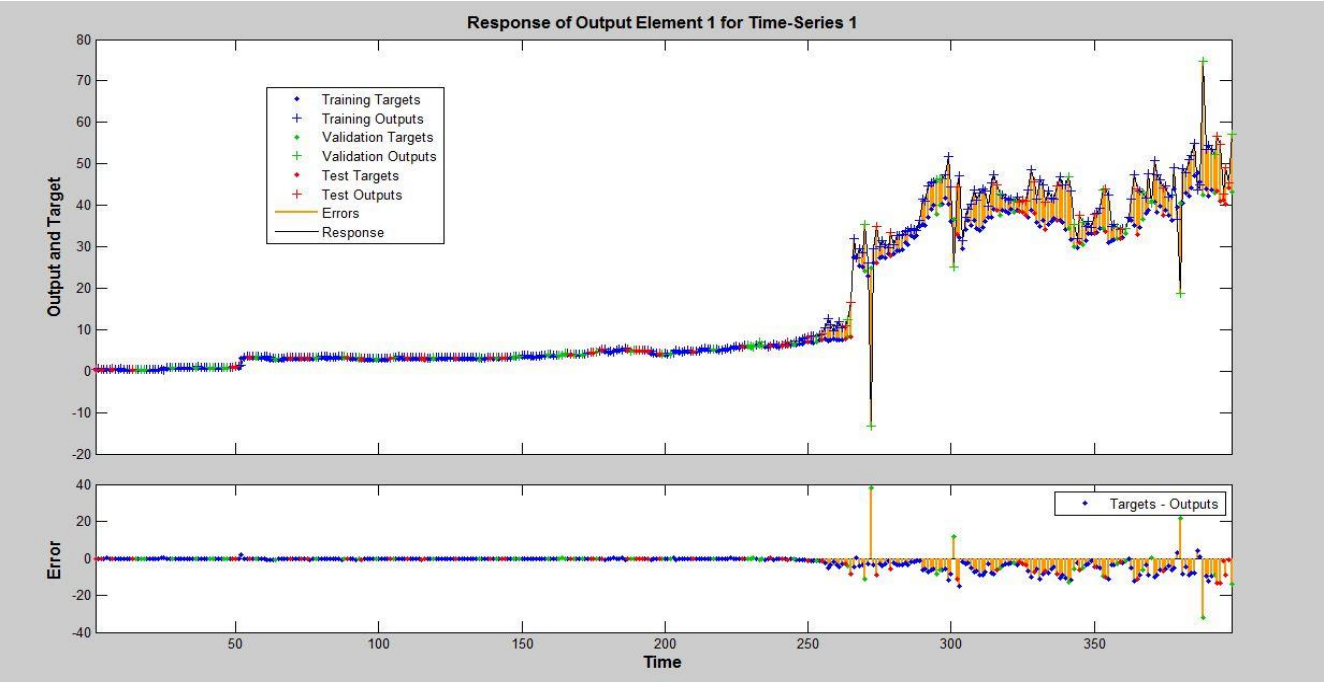


Future Prediction:

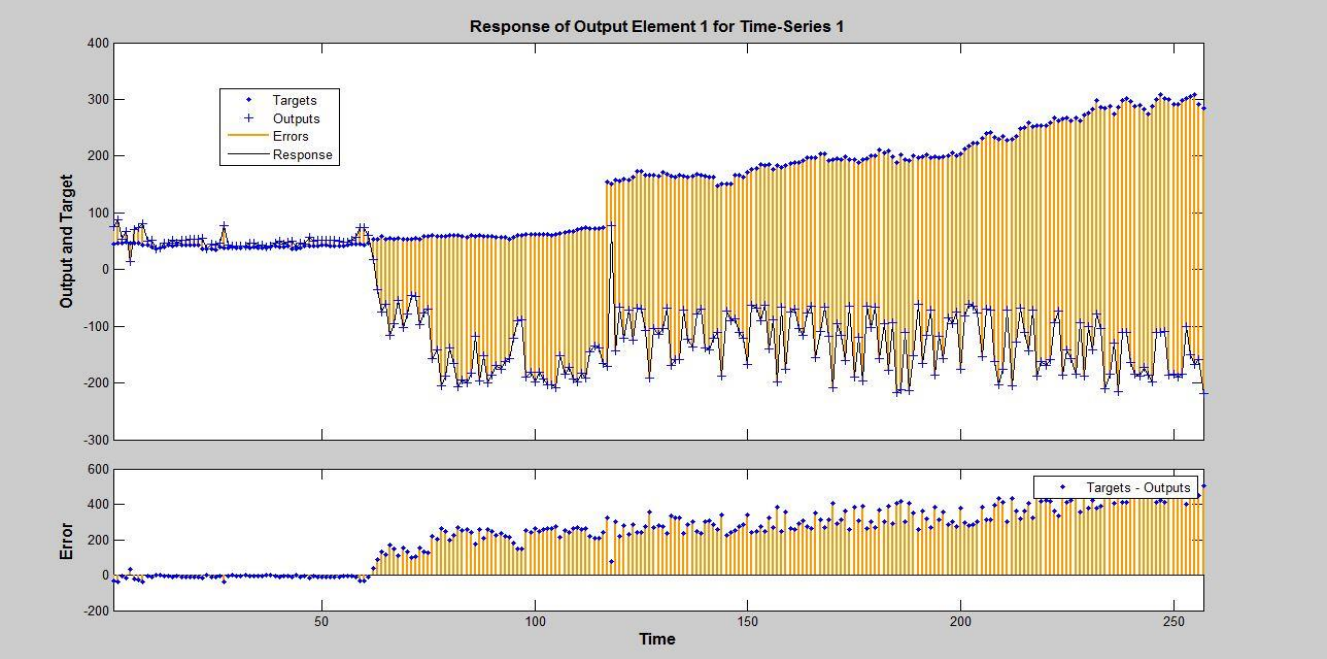


4. For a network with 200 hidden neurons and delay of 2:

Training data:



Future Prediction:



CONCLUSION FROM THE ABOVE RESULTS:

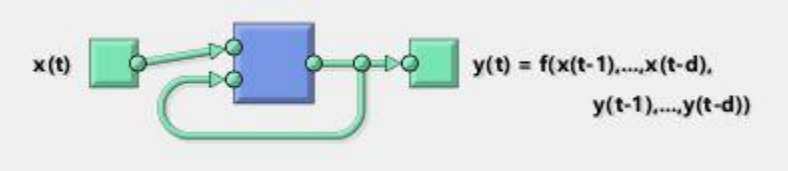
After performing the same test for different time series of stock prices we learn that the predictions show large deviations from the observed values after a relatively small number of time steps. Thus considering the chaotic nature of the time series of stock prices, prediction with an acceptable error can only be done upto a few time steps forward.

The fact that predictions for a longer period not working is not a minus of using neural networks over other methods but tells us about the chaotic nature of stock prices, and better results would be possible with a much more complicated model to estimate this time series.

And also we can see that as the number of neurons increased, the system performed better on training but failed to perform well on the future test set. This could be attributed to the inclination of the network to memorize the training data (network loses the ability to generalize).Hence smaller sized networks performed better on the future test data.

7. FUTURE WORK:

To better capture the chaotic nature of the time series of stock prices,a much more complicated model which is a combination of the above two methods known of NARX (Nonlinear AutoRegressive with eXogenous input) can be used.



In this method we could use another similar stock modeled as a time series, along with the data of historical prices of the same stock.

8. DATASET AND SOURCE CODE:

The dataset used in the entire project has been obtained from Yahoo Finance's Historical Data section.(http://in.finance.yahoo.com/q/hp?s=ITC.BO).

And the Source code for the basic neural network part was obtained from 'Matlab Neural Network Toolbox' (http://www.mathworks.in/products/neural-network/index.html)

9. REFERENCES:

[1] “On a Model for predicting the exchange rate Euro-Leu with A NAR neural network”, Published by Dumitru Ciobanu.

[2] “Machine Learning in Statistical Arbitrage” published by Xing Fu, Avinash Patra. (December,2009)

[3] “A Statistical Arbitrage Strategy” a master thesis project by Kun Zhu, Royal Institute of Technology, Stockholm, Sweden. (2005)

[4]Article on Arbitrage on Wikipedia

http://en.wikipedia.org/wiki/Arbitrage

[5] “Statistical arbitrage and tests of market efficiency” published by S.Horgan, R.Jarrow, and M. Warachka (2002).

[6] “An Improved test for Statistical arbitrage” published by Robert Jarrow, Melvyn Teo, Yiu Kuen Tse, Mitch Warachka (2011).

[7] “Statistical Arbitrage Stock Trading using Time Delay Neural Networks” a Machine learning final year project by Chris Pennock (Fall 2004).