Analysis of Statistical Arbitrage using Machine Learning Techniques in Stock Markets

AUSS Pradeep(10001)
Soren Goyal(10724)

What is Arbitrage?

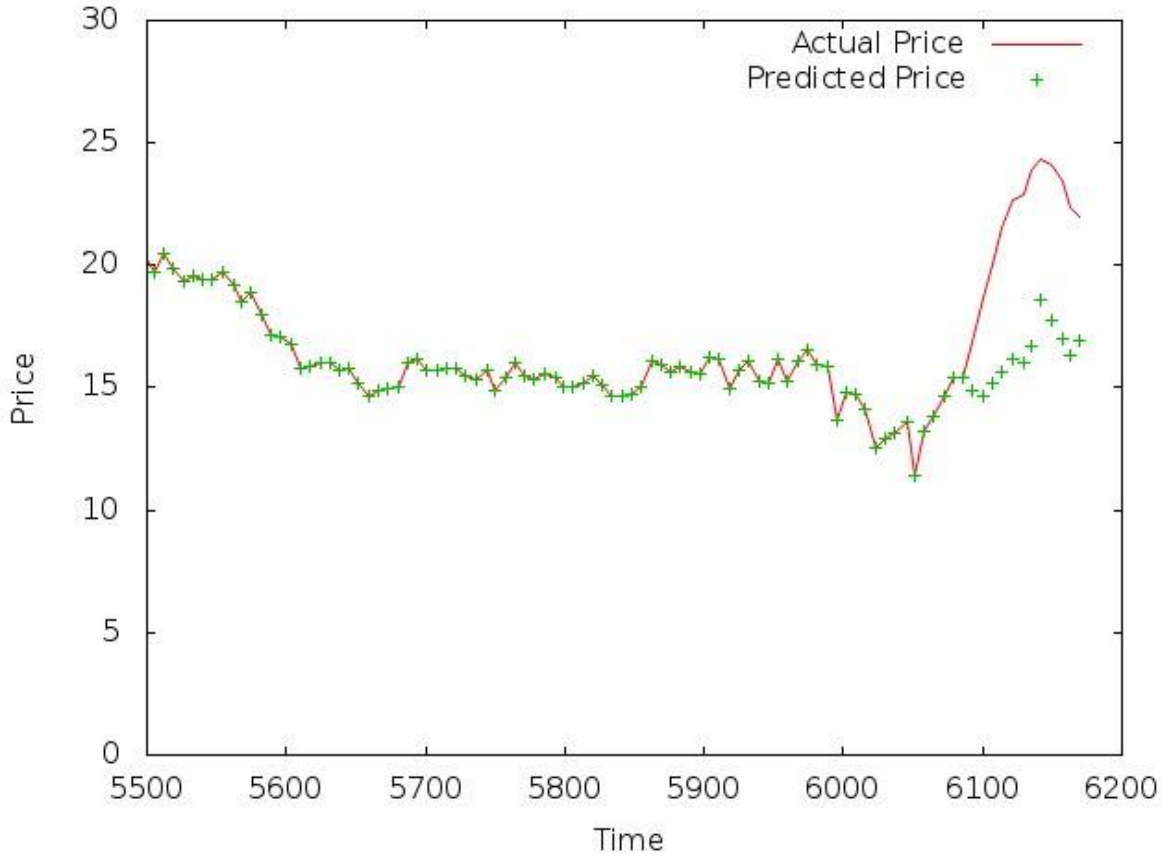
- **Buy the Doll in Kanpur and sell it in New Delhi**

Reason for Arbitrage

- **Market Inefficiencies** - In real world market information cannot be communicated without a time lag. This leads to different pricing of the same product in different markets.

Arbitrage in Stock Market

- The stock may be mispriced



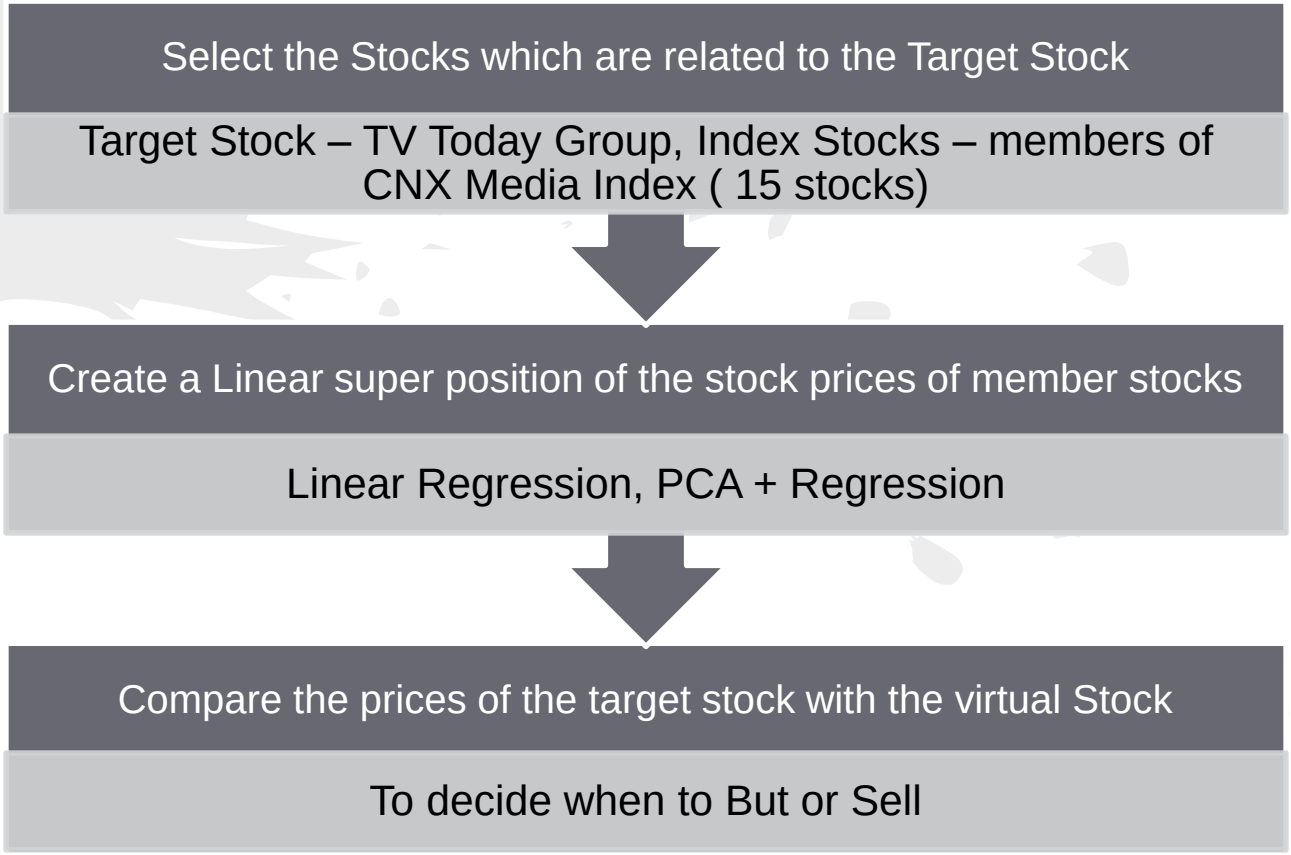
The Project

Predicting the price of the Stock based on historical Data.

- Virtual Stock
- Target Stock

Virtual Stock Method

- Create a virtual stock which mirrors the behavior of the Target stock.



Target Stock

Stocks used to create the "Virtual Stock"

JagranPrakashan	Ashtavinayak
Prime Focus	DB Corp
Reliance Media	DeccanChronicle
Sun TV Network	Den Networks
TV18 Broadcast	Dish TV India
Zee Entertain	Ent Network Ind
	Hathway Cable
	HT Media

CNX MEDIA

- Using regression we calculate the coefficient θ_i

$$P_t = \theta_0 + \sum_{i=1}^{100} \theta_i Q_{it}.$$

- The index created will represent the actual price of the Target Stock.
- If the price of the Target price is greater or lesser than the index, then the Stock is mispriced i.e, an Arbitrage has occurred.

Target Stock Method

- The data of the Target Stock is used to alone to generate predictions
- Method use – Fitting using TDNN

What is TDNN?

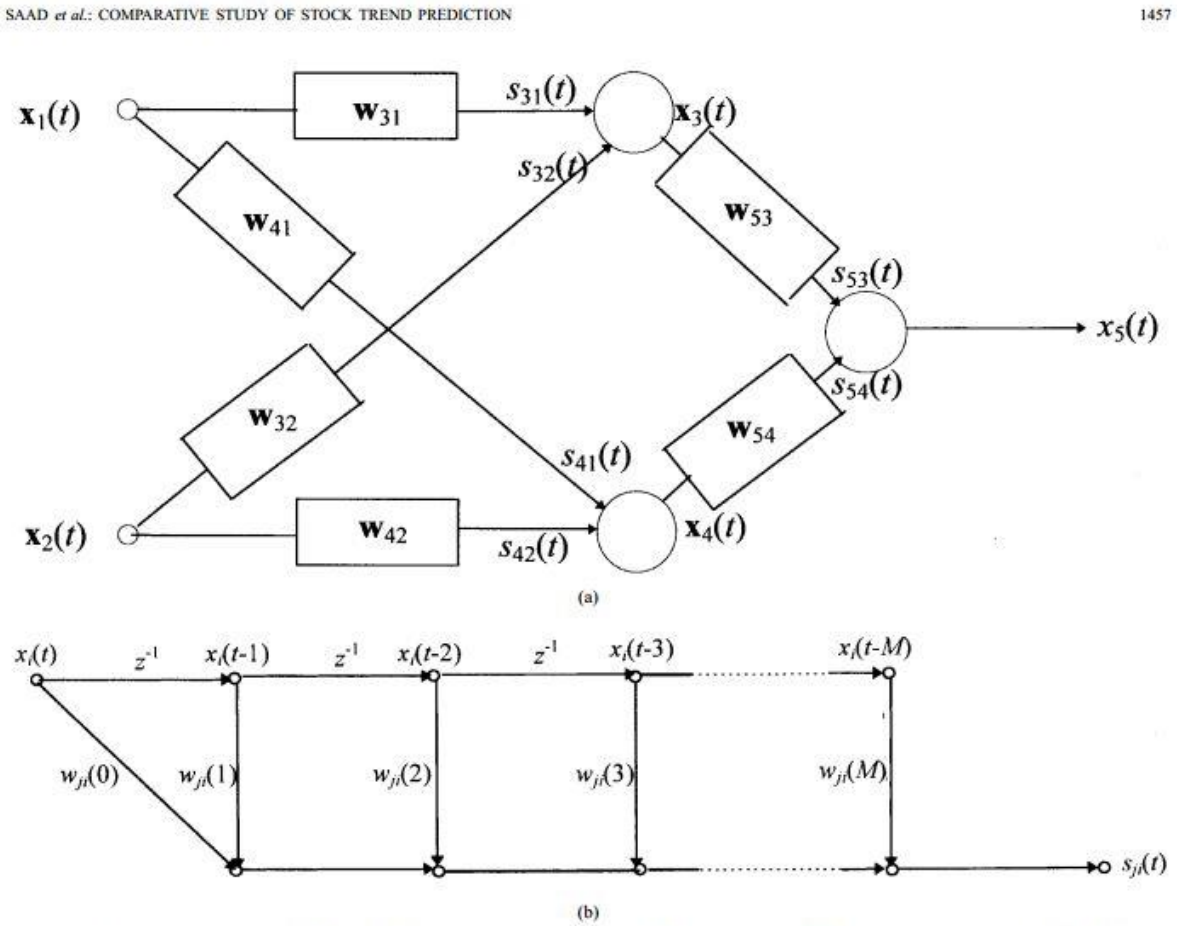


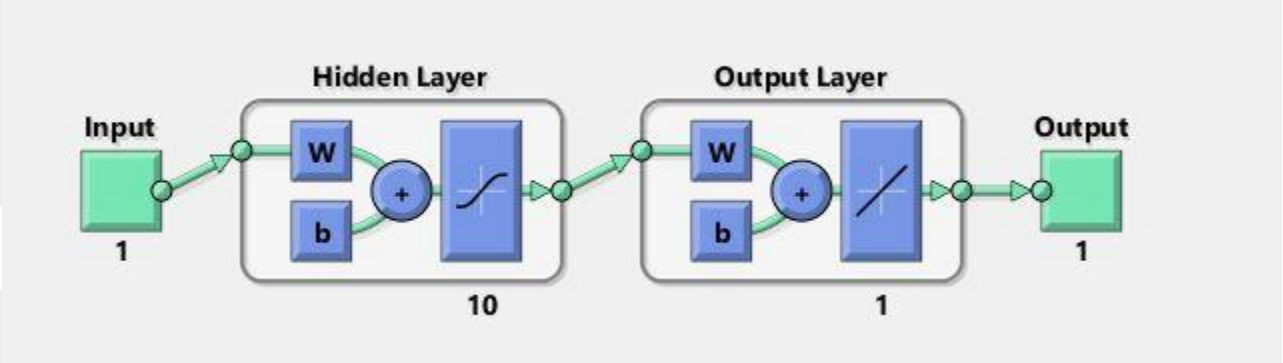
Fig. 1. (a) Three-neuron TDNN with FIR filters ($\mathbf{w}_{ji}$) as synaptic connections. (b) Expanded view of FIR synaptic connections of TDNN. FIR filters build internal memory into the network.

Source : Comparative Study of Stock Trend Prediction Using Time Delay, Recurrent and Probabilistic Neural Networks, 1998

Can a neural network be used to profit from the stock market?

Model of Neural Network

- 10 hidden layers with 10 neurons per layer.



- The data chosen varied in length, and because it was thought that these factors might affect the types of patterns contained in them and hence the performance of a learning machine.
- The length of each segment was either 100 and 900 weeks

Yahoo! Inc. (YHOO) - NasdaqGS

Add to Portfolio | Like 1.4K

23.48 ↓0.05(0.23%) 3:32PM EDT - Nasdaq Real Time Price

Historical Prices

Get Historical Prices for: GO

Set Date Range

Start Date: Apr 12 1996 Eg. Jan 1, 2010

End Date: Apr 1 2013

(•) Daily ( ) Weekly ( ) Monthly ( ) Dividends Only

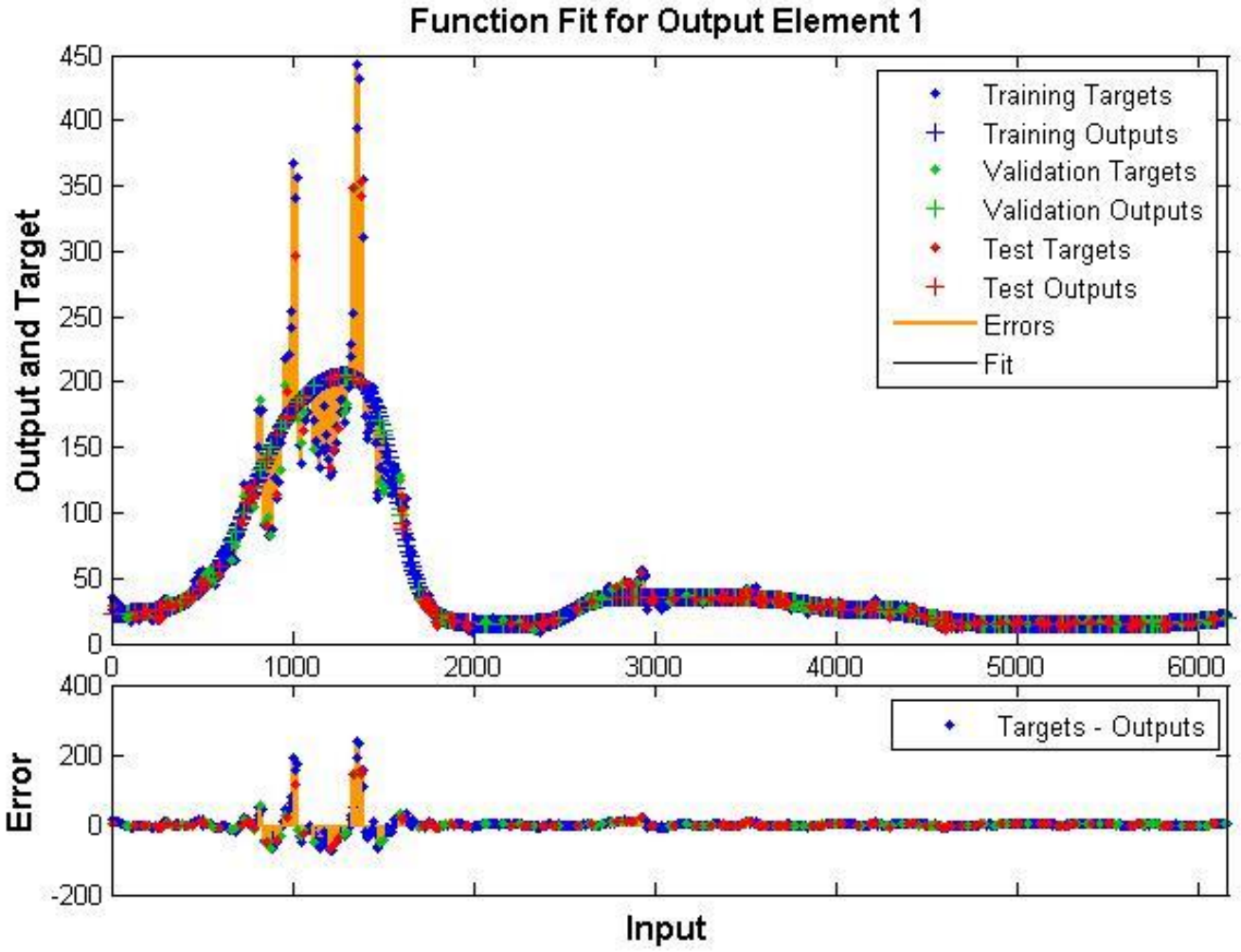
Get Prices

First | Previous | Next | Last

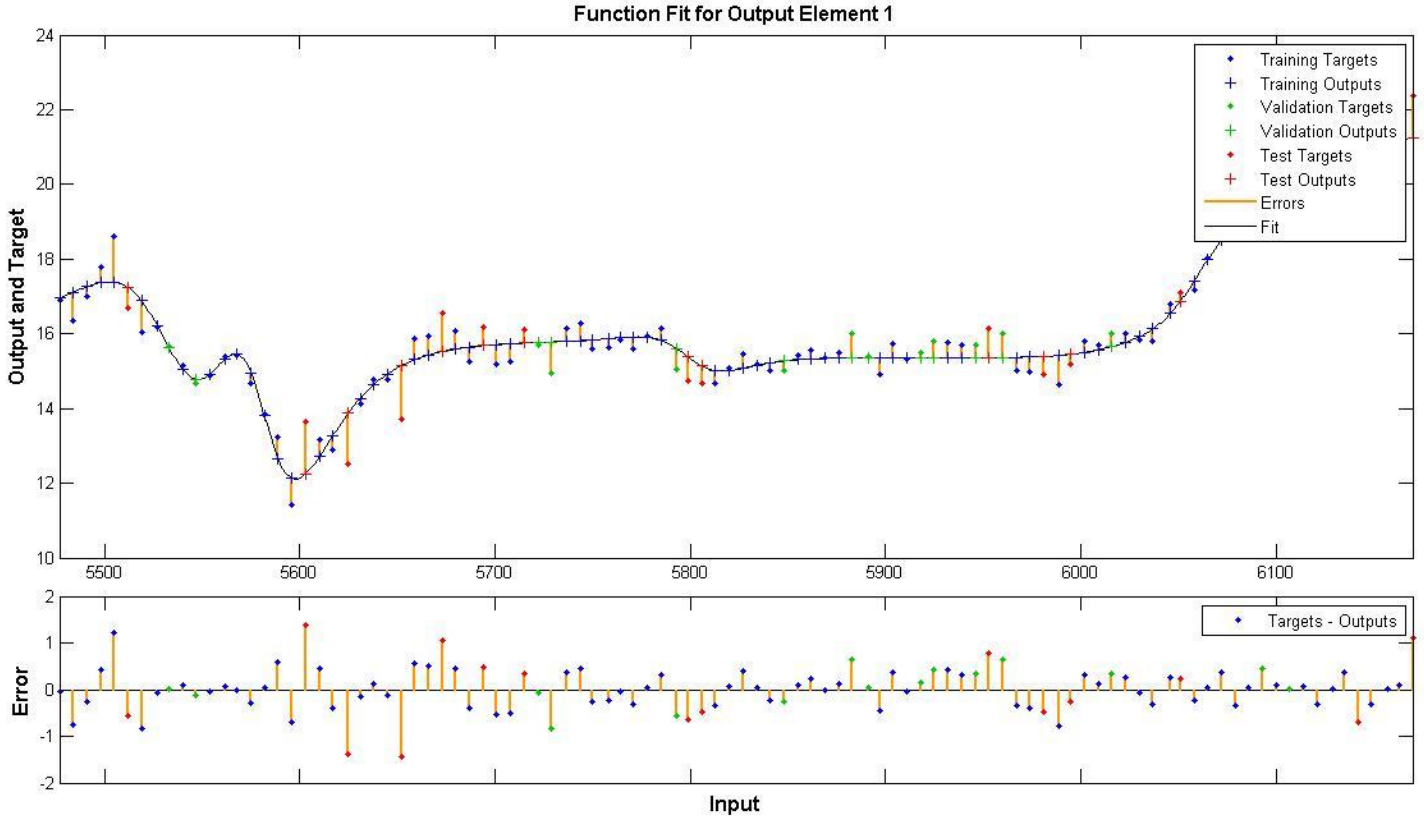
Prices

Date	Open	High	Low	Close	Volume	Adj Close*
Mar 28, 2013	23.63	23.77	23.45	23.53	17,611,900	23.53
Mar 27, 2013	23.54	23.83	23.41	23.59	13,943,600	23.59
Mar 26, 2013	23.46	23.62	23.35	23.59	16,878,400	23.59
Mar 25, 2013	23.41	23.88	23.31	23.38	23,138,900	23.38
Mar 22, 2013	22.88	23.26	22.72	23.26	17,956,100	23.26
Mar 21, 2013	22.39	22.95	22.36	22.86	24,719,100	22.86
Mar 20, 2013	22.03	22.33	21.95	22.10	18,094,100	22.10
Mar 19, 2013	22.06	22.33	21.89	22.17	12,846,900	22.17
Mar 18, 2013	21.90	22.17	21.87	22.01	15,071,700	22.01
Mar 15, 2013	22.34	22.39	21.97	22.07	33,557,400	22.07
Mar 14, 2013	22.47	22.75	22.41	22.43	12,798,500	22.43
Mar 13, 2013	22.48	22.48	22.16	22.34	13,956,200	22.34
Mar 12, 2013	22.51	22.58	22.19	22.40	12,012,300	22.40
Mar 11, 2013	22.80	23.00	22.57	22.60	16,489,200	22.60
Mar 8, 2013	22.92	22.96	22.71	22.90	10,583,500	22.90
Mar 7, 2013	22.92	23.00	22.65	22.70	12,881,800	22.70

Results


Function Fit for Output Element 1
Output and Target
Training Targets
Training Outputs
Validation Targets
Validation Outputs
Test Targets
Test Outputs
Errors
Fit
Error
Targets - Outputs
Input

- It was thought that a longer segment might simply provide more training data, or
- Conversely that a shorter segment might contain patterns more relevant to the near future movement of the stock.


Function Fit for Output Element 1
Output and Target
Training Targets
Training Outputs
Validation Targets
Validation Outputs
Test Targets
Test Outputs
Errors
Fit
24
22
20
18
16
14
12
10
5500
5600
5700
5800
5900
6000
6100
Error
Targets - Outputs
2
1
0
-1
-2
Input

- The raw closing price data was calculated at the end of a week. For example,for the training phase 75% of data was used, 15% for validation and 15% for testing.

- The trading strategy learned by the networks was as follows: If at the end of a sequence of 50 daily closing prices, the price will go up by 2% or more over the next 20 days, buy it. This strategy was chosen as a balance between two opposing forces.
- On one hand, it is easier to predict the movement of a time series a short period into the future.

- These trends are promising, and would seem to indicate that by training a machine on the proper amount of data and by using the right type of learning machine, one might be able to beat the performance of an individual stock, on average

Status

- Out of the 3 mentioned methods, we have successfully implemented the method involving TDNN.
- Beyond this we will compare the performance of the methods on the same stock.
- Further, we will like to include “Recurrent and Probabilistic Neural Networks”

References

- [1] “Machine Learning in Statistical Arbitrage” published by Xing Fu, Avinash Patra. (December,2009)
- [2] “A Statistical Arbitrage Strategy” a master thesis project by Kun Zhu, Royal Institute of Technology, Stockholm, Sweden. (2005)
- [3]Article on Arbitrage on Wikipedia http://en.wikipedia.org/wiki/Arbitrage
- [4] “Statistical arbitrage and tests of market efficiency” published by S.Horgan, R.Jarrow, and M. Warachka (2002).
- [5] “An Improved test for Statistical arbitrage” published by Robert Jarrow, Melvyn Teo, Yiu Kuen Tse, Mitch Warachka (2011).
- [6] “Statistical Arbitrage Stock Trading using Time Delay Neural Networks” a Machine learning final year project by Chris Pennock (Fall 2004).

Three Kinds of Samples:

Training:

These are presented to the network during training, and the network is adjusted according to its error.

Validation:

These are used to measure network generalization, and to halt training when generalization stops improving.

Testing:

These have no effect on training and so provide an independent measure of network performance during and after training.