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.

Select the Stocks which are related to the Target Stock

Target Stock – TV Today Group, Index Stocks – members of CNX Media Index (15 stocks)

Create a Linear super position of the stock prices of member stocks

Linear Regression, PCA + Regression

Compare the prices of the target stock with the virtual Stock

To decide when to But or Sell
Stocks used to create the “Virtual Stock”

<table>
<thead>
<tr>
<th>JagranPrakashan</th>
<th>Ashtavinayak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prime Focus</td>
<td>DB Corp</td>
</tr>
<tr>
<td>Reliance Media</td>
<td>DeccanChronicle</td>
</tr>
<tr>
<td>Sun TV Network</td>
<td>Den Networks</td>
</tr>
<tr>
<td>TV18 Broadcast</td>
<td>Dish TV India</td>
</tr>
<tr>
<td>Zee Entertain</td>
<td>Ent Network Ind</td>
</tr>
<tr>
<td></td>
<td>Hathway Cable</td>
</tr>
<tr>
<td></td>
<td>HT Media</td>
</tr>
</tbody>
</table>

Target Stock

CNX MEDIA
Using regression we calculate the coefficient $\theta_i$.

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

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

Fig. 1. (a) Three-neuron TDNN with FIR filters ($w_{ij}$) as synaptic connections. (b) Expanded view of FIR synaptic connections of TDNN. FIR filters build internal memory into the network.
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
### Historical Prices

**Yahoo! Inc. (YHOO) - NasdaqGS**

<table>
<thead>
<tr>
<th>Date</th>
<th>Open</th>
<th>High</th>
<th>Low</th>
<th>Close</th>
<th>Volume</th>
<th>Adj Close*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar 28, 2013</td>
<td>23.83</td>
<td>23.77</td>
<td>23.45</td>
<td>23.53</td>
<td>17,611,900</td>
<td>23.53</td>
</tr>
<tr>
<td>Mar 27, 2013</td>
<td>23.54</td>
<td>23.83</td>
<td>23.41</td>
<td>23.59</td>
<td>13,943,500</td>
<td>23.59</td>
</tr>
<tr>
<td>Mar 26, 2013</td>
<td>23.46</td>
<td>23.62</td>
<td>23.35</td>
<td>23.59</td>
<td>16,878,400</td>
<td>23.59</td>
</tr>
<tr>
<td>Mar 22, 2013</td>
<td>22.88</td>
<td>23.26</td>
<td>22.72</td>
<td>23.26</td>
<td>17,956,100</td>
<td>23.26</td>
</tr>
<tr>
<td>Mar 21, 2013</td>
<td>22.39</td>
<td>22.96</td>
<td>22.38</td>
<td>22.86</td>
<td>24,719,100</td>
<td>22.86</td>
</tr>
<tr>
<td>Mar 20, 2013</td>
<td>22.03</td>
<td>22.33</td>
<td>21.95</td>
<td>22.10</td>
<td>18,094,100</td>
<td>22.10</td>
</tr>
<tr>
<td>Mar 19, 2013</td>
<td>22.06</td>
<td>22.33</td>
<td>21.89</td>
<td>22.17</td>
<td>12,846,900</td>
<td>22.17</td>
</tr>
<tr>
<td>Mar 18, 2013</td>
<td>21.90</td>
<td>22.17</td>
<td>21.87</td>
<td>22.01</td>
<td>15,071,700</td>
<td>22.01</td>
</tr>
<tr>
<td>Mar 15, 2013</td>
<td>22.34</td>
<td>22.39</td>
<td>21.97</td>
<td>22.07</td>
<td>33,557,400</td>
<td>22.07</td>
</tr>
<tr>
<td>Mar 14, 2013</td>
<td>22.47</td>
<td>22.75</td>
<td>22.41</td>
<td>22.43</td>
<td>12,798,500</td>
<td>22.43</td>
</tr>
<tr>
<td>Mar 13, 2013</td>
<td>22.48</td>
<td>22.48</td>
<td>22.16</td>
<td>22.34</td>
<td>13,856,200</td>
<td>22.34</td>
</tr>
<tr>
<td>Mar 12, 2013</td>
<td>22.51</td>
<td>22.56</td>
<td>22.19</td>
<td>22.40</td>
<td>12,012,300</td>
<td>22.40</td>
</tr>
<tr>
<td>Mar 11, 2013</td>
<td>22.80</td>
<td>23.00</td>
<td>22.57</td>
<td>22.60</td>
<td>16,489,200</td>
<td>22.60</td>
</tr>
<tr>
<td>Mar 8, 2013</td>
<td>22.92</td>
<td>22.96</td>
<td>22.71</td>
<td>22.90</td>
<td>10,583,500</td>
<td>22.90</td>
</tr>
<tr>
<td>Mar 7, 2013</td>
<td>22.92</td>
<td>23.00</td>
<td>22.65</td>
<td>22.70</td>
<td>12,881,800</td>
<td>22.70</td>
</tr>
</tbody>
</table>
Results

Function Fit for Output Element 1

![Function Fit Plot](image)

- Training Targets
- Training Outputs
- Validation Targets
- Validation Outputs
- Test Targets
- Test Outputs
- Errors
- Fit

**Output and Target**

**Error**

**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

Error

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

• [3] Article on Arbitrage on Wikipedia [link]
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.