Chapter 4

Stock Market Prediction from Sectoral Indices using an Adaptive Network Based Fuzzy Inference System

4.1 Introduction

The effect of different variables on the stock market is studied in this Chapter. Forecasting of stock market returns is difficult because market volatility needs to be captured in more reliable and user-friendly models. Among the other factors, accurate modeling needs, consideration of the phenomenon that can explain market volatility more precisely. With their professional knowledge, financial analysts and stock fund managers predict stock market by using stock analyzing tools based on technical analysis by Chuang [32], fundamental analysis by Atsalaki [7] and time series models by Box [16].

In finance, stock market prediction and choice of predictive variables are very important topics. Researchers, stock buyers, stock sellers and many more who are involved in the financial sector are always in search of new techniques for stock market prediction. Therefore, enormous concepts and techniques are demonstrated for forecasting stock prices. The empirical study on tactical asset allocation and
forecasting has been done by many researchers (Faber [41], Jensen et al. [59], Rastogi et al. [84], rastogi and Dhar [83]).

To further improve the forecasting accuracy, a new fuzzy time series model was proposed by Chen and Chung [26, 27] which used genetic algorithms to adjust the length of linguistic intervals. To demonstrate fuzzy non-linear relationships in the fuzzy time-series model (Huarng and Yu [54]) employed a neural network. Chen et al. [28] demonstrated that the process of mining fuzzy logical relationships is not easily understandable. Moreover, all the models that are mentioned above have been limited to only one variable application (Yu [98]). Chen et al. [28] suggested a comprehensive fuzzy time-series, which factors in recent periods of stock prices and fuzzy logical relationships into the prediction procedure. Huarng [98] proposed a bivariate model, which applies neural networks to fuzzy time-series forecasting.

We have found a major drawback in the existing literature of financial forecasting, which is listed below:

- Nearly all conventional time-series models use only one variable in stock market forecasting, although lots of noise is caused by changes in the market conditions (Yu [98]).

To overcome this drawback, we study the effect of different sectors (variables) of National Stock Exchange (NSE) on CNX S&P Nifty index (benchmark index of NSE).

4.2 Methodology and Tools

Stock markets can be divided into thematic sectors based on a company’s area of operation to which the stock is linked. The number of sectors depends upon market diversification, type of market and the number of companies listed on the stock market. Prices of individual sectors have the greater tendency to move along the same direction as the overall stock market. This is specifically true if the sector is among the largest or most heavily traded on the stock market. So, we have used sectors as variables to forecast Indian stock market.
The sectors which affect the CNX S&P Nifty index to a greater extent are chosen by stepwise regression analysis (SRA). To minimize the complexity of the model, data dimensions are reduced with the help of principal component analysis (PCA). Furthermore, a hybrid model is proposed in which fuzzy c-means (FCM) clustering method and adaptive neuro-fuzzy inference system (ANFIS) are used for fuzzification and for defining fuzzy relations. This fuzzy inference system can model the useful expressions of human knowledge, which can be very helpful for stock market investors. The schematic diagram of proposed model is shown in Figure 4.1.

4.2.1 Forward Selection of Attributes

SRA is used to choose an optimal set of variables from the set of potentially useful variables. To select the important factors in their systems, researchers have applied a variety of techniques such as grey relation analysis by Chang [23], genetic algorithms by Elalami [37] and stepwise regression analysis by Chang [22]. In recent years researchers have applied SRA for the selection of input variables in the field of stock market forecasting (Chang and Liu [21], Hadavandi et al. [51]). In this study, we used SRA for the selection of input variables to reduce the complexity of the forecasting model and to improve the forecasting accuracy. SRA gives the set of independent variables that are most closely related to the dependent variables. In it’s procedure, a method for adding variables or removing variables is used to find the most primed combination of sectors for forecasting stock prices. For every step, simple regression analysis is performed using the previously added independent variables and one of the removed variables.

4.2.2 Dimensionality Reduction

Basically, PCA is used to maximize the correlation between the original variables to form a new set of variables that are mutually uncorrelated (Jolliffe [61]). A basic assumption in the use of PCA is that loading and score vectors representing the largest eigenvalue hold the most useful information related to the particular problem and the rest mainly comprise noise. This means that the first principal component explains most of the variance in the data, and each subsequent one accounts for the
largest proportion of variability that has not been accounted for by its predecessors. PCA has been employed in many studies (Zhang and Vaidya [88, 106]).

4.3 Experimental Results

In this section, we have presented experimental results. Results are mainly divided into four parts which are, data collection, selection of input variables by SRA and dimensionality reduction by PCA and finally generating a forecasting model. The step-wise description of each part is given below:

4.3.1 Collection of Data Sets

Data has been collected from different sectors such as Auto, Bank, Energy, Finance, FMCG, IT, Media, Metal, Pharma, PSU bank, Realty and Nifty S&P index of NSE for the period Jan 1, 2008 to Jan 31, 2012. Data is divided into two sets, first set Jan 1, 2008 to Dec 31, 2010 has been used as training data and second set from Jan 1, 2011 to Jan 31, 2012 is used to test the forecasting model. The data contains eleven independent variables and one i.e. Nifty S&P dependent variable.

4.3.2 Selection of Input Variables by SRA

In the first stage, we have used SRA to remove low impact sectors and choose the most significant ones. The standard criterion for the addition and removal of a variable is determined by $F$-test and to decrease the sum of squared error values. After entering the first variable, the number of variables are increased step by step in the model. Once a variable is removed from the model, that will never enter the model again. Firstly, the values $F$-to-enter ($F_e$) and $F$-to-remove ($F_r$) are calculated. Then the $F$-value for each step is determined. If $F > F_e$, then the variable is considered to enter in the model and if $F < F_r$, the variable is removed from the model. We used the values of $F_e$ and $F_r$ as $F_e \leq 0.050$ and $F_r \geq 0.100$. We performed SRA with the help of SPSS software. The results given by SPSS are shown in Table 4.1, the outcomes from this analysis are Finance, Energy, Pharma, Realty, IT and PSU Bank.
Table 4.1: SRA Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables Entered</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Finance</td>
<td>Stepwise (Criteria: $F_e \leq 0.050$, $F_r \geq 0.100$)</td>
<td></td>
</tr>
<tr>
<td>2 Energy</td>
<td>Stepwise (Criteria: $F_e \leq 0.050$, $F_r \geq 0.100$)</td>
<td></td>
</tr>
<tr>
<td>3 Pharma</td>
<td>Stepwise (Criteria: $F_e \leq 0.050$, $F_r \geq 0.100$)</td>
<td></td>
</tr>
<tr>
<td>4 Realty</td>
<td>Stepwise (Criteria: $F_e \leq 0.050$, $F_r \geq 0.100$)</td>
<td></td>
</tr>
<tr>
<td>5 IT</td>
<td>Stepwise (Criteria: $F_e \leq 0.050$, $F_r \geq 0.100$)</td>
<td></td>
</tr>
<tr>
<td>6 PSU Bank</td>
<td>Stepwise (Criteria: $F_e \leq 0.050$, $F_r \geq 0.100$)</td>
<td></td>
</tr>
</tbody>
</table>

4.3.3 Reduction of the Data Dimensions by PCA

Let we have data sample named $X_{n \times m}$, where $n$ is the number of data samples and $m$ is number of variables. We have six variables which are chosen with the help of SRA, so value of $m = 6$. We performed PCA with the help of SPSS software.

Firstly, the data set should be standardized to reduce the effect of dimensions and grade between each variable, this can be done by using the equation

$$y_{ij} = \frac{x_{ij} - (1/n) \sum_{i=1}^{n} x_{ij}}{\sqrt{(1/n - 1) \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}},$$ \hspace{1cm} (4.3.1)

where $y_{ij}$ is the standardized value, $x_{ij}$ is the data point of $i$-th row and $j$-th column, $\bar{x}_j$ is the mean value of $j$-th column and $n$ is the number of total data points. Then correlation coefficient matrix $M$ is obtained according to $y_{ij}$.

$$M = \frac{1}{n-1} YY^\prime,$$ \hspace{1cm} (4.3.2)

where $Y$ is the standardize data matrix and $Y^\prime$ is transpose of $Y$. Table 4.2 gives the coefficient of correlation for each variable.

Now eigenvalues $\lambda_j$ and eigenvector $E_j$ of the matrix $M$ are calculated. The vari-
Table 4.2: Correlation coefficient matrix of selected attributes

<table>
<thead>
<tr>
<th></th>
<th>Finance</th>
<th>Energy</th>
<th>Pharma</th>
<th>Realty</th>
<th>IT</th>
<th>PSU Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finance</td>
<td>1</td>
<td>0.785</td>
<td>0.860</td>
<td>0.075</td>
<td>0.912</td>
<td>0.948</td>
</tr>
<tr>
<td>Energy</td>
<td>0.785</td>
<td>1</td>
<td>0.496</td>
<td>0.495</td>
<td>0.645</td>
<td>0.709</td>
</tr>
<tr>
<td>Pharma</td>
<td>0.860</td>
<td>0.496</td>
<td>1</td>
<td>-0.271</td>
<td>0.946</td>
<td>0.816</td>
</tr>
<tr>
<td>Realty</td>
<td>0.075</td>
<td>0.495</td>
<td>-0.271</td>
<td>1</td>
<td>-0.169</td>
<td>-0.082</td>
</tr>
<tr>
<td>IT</td>
<td>0.912</td>
<td>0.645</td>
<td>0.946</td>
<td>-0.169</td>
<td>1</td>
<td>0.887</td>
</tr>
<tr>
<td>PSU Bank</td>
<td>0.948</td>
<td>0.709</td>
<td>0.816</td>
<td>-0.082</td>
<td>0.887</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.3: Eigenvalues and cumulative variance

<table>
<thead>
<tr>
<th># of components</th>
<th>Eigenvalues</th>
<th>% of Variance</th>
<th>cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.227</td>
<td>70.445</td>
<td>70.445</td>
</tr>
<tr>
<td>2</td>
<td>1.415</td>
<td>23.589</td>
<td>94.034</td>
</tr>
<tr>
<td>3</td>
<td>0.193</td>
<td>3.211</td>
<td>97.245</td>
</tr>
<tr>
<td>4</td>
<td>0.113</td>
<td>1.886</td>
<td>99.131</td>
</tr>
<tr>
<td>5</td>
<td>0.034</td>
<td>0.573</td>
<td>99.704</td>
</tr>
<tr>
<td>6</td>
<td>0.018</td>
<td>0.296</td>
<td>100.000</td>
</tr>
</tbody>
</table>

The variance and cumulative variance of each principal component is calculated by following equations:

\[
\eta_j = \frac{\lambda_j}{\sum_{j=1}^{m} \lambda_j}, \quad (4.3.3)
\]

\[
\eta_k = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{m} \lambda_j}, \quad (4.3.4)
\]

where \( \eta_j \) is the variance of \( j \)-th principal component, \( \eta_k \) is the cumulative variance of \( k \)-th principal component, \( m \) is the number of variables in the standardized data set. The eigenvalues, variance and cumulative variance is given in the Table 4.3.

We fix the variance rate as greater than one for each principal component. So first four components are taken to explain the information contained by original data. The loading value \( \alpha_j \), (Table 4.4) for selected principal components is calculated by the equation (4.3.5).

\[
\alpha_j = \sum_{j=1}^{m} \sqrt{\lambda_j E_j}. \quad (4.3.5)
\]
Table 4.4: Loading values $\alpha$ for selected principal components

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finance</td>
<td>0.983</td>
<td>0.078</td>
<td>-0.017</td>
<td>0.123</td>
</tr>
<tr>
<td>Energy</td>
<td>0.779</td>
<td>0.566</td>
<td>-0.137</td>
<td>-0.229</td>
</tr>
<tr>
<td>Pharma</td>
<td>0.906</td>
<td>-0.315</td>
<td>0.266</td>
<td>-0.004</td>
</tr>
<tr>
<td>Realty</td>
<td>-0.009</td>
<td>0.977</td>
<td>0.175</td>
<td>0.114</td>
</tr>
<tr>
<td>IT</td>
<td>0.962</td>
<td>-0.180</td>
<td>0.119</td>
<td>-0.090</td>
</tr>
<tr>
<td>PSU Bank</td>
<td>0.954</td>
<td>-0.053</td>
<td>-0.241</td>
<td>0.156</td>
</tr>
</tbody>
</table>

The loading values give the related coefficients between variables and principal components. From the Table 4.4 we can see that the component C1 and C2 are related positively to Finance and Energy. Similarly, component C2 has a negative relation with pharma and so on. Finally, the data sets for each principal component are computed and a new data matrix for the input to forecast model is constructed as below:

\[
C1 = 0.983 \times Finance + 0.779 \times Energy + 0.906 \times Pharma - 0.009 \times Realty \\
+ 0.962 \times IT + 0.954 \times PSU Bank.
\]

\[
C2 = 0.078 \times Finance + 0.566 \times Energy - 0.315 \times Pharma + 0.977 \times Realty \\
- 0.180 \times IT - 0.053 \times PSU Bank.
\]

\[
C3 = -0.017 \times Finance - 0.137 \times Energy - 0.266 \times Pharma + 0.175 \times Realty \\
+ 0.119 \times IT - 0.241 \times PSU Bank.
\]

\[
C4 = 0.123 \times Finance - 0.229 \times Energy - 0.004 \times Pharma + 0.114 \times Realty \\
- 0.090 \times IT + 0.156 \times PSU Bank.
\]

We performed error calculations with two and four principal components for prediction. It is observed that with two principal component the value of RMSE is 154 and with four RMSE becomes 151 (Table 4.5). Hence, there is no significant difference between RMSE values calculated using two and four components. So to reduce the complexity of the model in the rest of the analysis we have chosen two principal components i.e. C1 and C2 with minimum RMSE value for the construction of input data matrix for forecasting.
Table 4.5: RMSE value for different combinations of principal components

<table>
<thead>
<tr>
<th>Principal Component combination</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1, C2, C3 and C4</td>
<td>151</td>
</tr>
<tr>
<td>C1 and C2</td>
<td>154</td>
</tr>
<tr>
<td>C2 and C3</td>
<td>156</td>
</tr>
<tr>
<td>C3 and C4</td>
<td>154.5</td>
</tr>
<tr>
<td>C4 and C1</td>
<td>155</td>
</tr>
</tbody>
</table>

### 4.3.4 Forecasting Model

**Step 1:** Partitioning of data set to extract linguistic rules:

Firstly, the input variables determined by PCA, $C_1(t)$ and $C_2(t)$ are divided into three clusters with the help of FCM clustering (section 2.7.1). The parameters of FCM clustering are given below:

- maximum number of iterations = 100,
- exponent for the partition matrix = 2.0,
- minimum amount of improvement = 0.007.

These clusters are used as linguistic intervals i.e. low (L), middle (M), and high (H). The Gaussian membership function is used as input membership function. For output, we set linear type of membership function. Now, Sugeno fuzzy model is employed to generate fuzzy 'if-then' rules, the general rule in the inference system is described as:

if $x(C_1(t)) = A_i, y(C_2(t)) = B_i$ then $f_i = p_ix + q_iy + r_i$.

Where $x(C_1(t)), y(C_2(t))$ are linguistic variable, $A_i, B_i$ are the linguistic values (high, middle, low), $f_i$ denotes the $i$-th output and $p_i, q_i, r_i$ are the parameters ($i = 1, 2, 3$).

The rules formed with three linguistic intervals (low (L), middle (M), and high (H)) are described as follows:
Table 4.6: Output membership function parameters

<table>
<thead>
<tr>
<th></th>
<th>$p_i$</th>
<th>$q_i$</th>
<th>$r_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>0.2013</td>
<td>-0.05031</td>
<td>575.9</td>
</tr>
<tr>
<td>middle</td>
<td>0.2044</td>
<td>0.2068</td>
<td>-120.1</td>
</tr>
<tr>
<td>high</td>
<td>0.1871</td>
<td>0.2317</td>
<td>226.3</td>
</tr>
</tbody>
</table>

Rule 1.
if $x(C1(t)) = A_L$, $y(C2(t)) = B_L$ then $f_L(Nifty \text{ S&P}_{t+1}) = a_L x + b_L y + r_L$.

Rule 2.
if $x(C1(t)) = A_M$, $y(C2(t)) = B_M$ then $f_M(Nifty \text{ S&P}_{t+1}) = a_M x + b_M y + r_M$.

Rule 3.
if $x(C1(t)) = A_H$, $y(C2(t)) = B_H$ then $f_H(Nifty \text{ S&P}_{t+1}) = a_H x + b_H y + r_H$.

Step 2: Training of parameters by fuzzy adaptive network:
In this section, a combination of the least-squares method and the back propagation gradient descent method is applied for training process to get optimal FIS parameters. ANFIS employs gradient descent to fine-tune premise parameters that define membership functions and for consequent parameters that define the coefficients of each output equation, ANFIS uses the least-squares method to identify them. In this study, we set the stopping criterion of 100 epoch and obtained optimal fuzzy inference parameters. The output membership function parameters for the extracted rules are shown in Table 4.6.

Step 3: The proposed model for forecasting the future price:
After obtaining the optimal membership functions, the data set is trained along with optimal fuzzy rules and membership functions to forecast the future price $f(t + 1)$ in the testing data sets. The forecasting results of proposed model are shown in Figure 4.2.

Step 4: Evaluation of forecasting performance:
To compare the performance of proposed model with the existing models, root mean square error (RMSE) is chosen as evaluation criterion defined as:

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\text{actual}(t_i) - \text{forecast}(t_i))^2},
\]

where \(\text{actual}(t_i)\) refers to the actual value of \(i^{th}\) data point, \(\text{forecast}(t_i)\) refers to the predicted value of \(i^{th}\) data point and \(N\) is the total number of data entries. The actual and predicted values are shown in Figure 4.2. Further, the proposed model is compared with some existing models and the results are shown in Table 4.7.

### 4.4 Discussion

The main aim of this work is to present a model to predict efficiently the stock market behavior. Successful prediction of stock prices may promise very attractive benefits, it effects a financial trader’s decision to buy or sell an element. There are a number of factors that influence the stock market, so these tasks are very complicated. Artificial intelligence techniques are successfully employed to solve the problems of stock markets. In this chapter, the variables which effect the stock market to a greater extent are chosen by applying stepwise regression analysis. Data dimensions are reduced with the help of principal component analysis. After ANFIS combined with FCM is then applied to predict stock prices of NSE index CNX S & P Nifty effectively. Further, the proposed model is compared with Autoregressive Integrated Moving Average Model (ARIMA) [16] and Huarng’s [54] model. The results show that the proposed model outperforms the listing models with a minimum value of RMSE (Table 4.7). The study reveals that the performance of stock market can be significantly predicted by the proposed model. The prediction of CNX S & P Nifty
from Jan 1, 2011 to Dec 31, 2012 is shown in Figure 4.2, from which, we can see that the actual value and predicted value are very close to each other in the testing data set. The proposed model achieves an accuracy of up to 80%.

4.5 Conclusion

An expert system to predict the stock prices was proposed here, which has used regression analysis and PCA to select the input variables. After the selection of input variables, FCM has been used for clustering the data set and adaptive neuro-fuzzy inference system for fuzzification and for defining fuzzy relations.

It is observed that the proposed method has advantages over existing models, such as problems caused by partitioning of the universe of discourse are removed with the use of FCM in the fuzzification step. Since fuzzy relationships are defined by artificial neural networks, so there is no need to use complex fuzzy group relationship tables or difficult matrix operations. Secondly, only three rules are generated, which reduced the computational complexity of the model. Finally, more than one variable is used for the prediction of stock market prices, because the prediction accuracy is increased with the use of more than one variable so the proposed system can better predict the actual market environment. Therefore, proposed hybrid fuzzy time series approach will be a good choice for forecasting of the stock market index.
Stock index
data sector wise

Choose significant sectors by regression

Apply PCA to reduce data dimensions

Generate forecasting model

Partition the data set with FCM

Create fuzzy rules

Train network

Required accuracy

No

Yes

Stock market forecasting

Figure 4.1: Schematic diagram of proposed model sequence
Figure 4.2: The forecasting results of proposed model