Stock Market Trend Prediction Using ARIMA-based Neural Networks

Jung-Hua Wang and Jia-Yann Leu
Department of Electrical Engineering
National Taiwan Ocean University
2, Pei-Ning Rd, Keelung, Taiwan
Phone : 886-2-462-2192 Ext. 6207
E-mail : jhwang@celab1.ee.ntou.edu.tw

Abstract

We develop a prediction system useful in forecasting mid-term price trend in Taiwan stock market (Taiwan stock exchange weighted stock index, abbreviated as TSEWSI). The system is based on a recurrent neural network trained by using features extracted from ARIMA analyses. By differencing the raw data of the TSEWSI series and then examining the ACF and PACF plots, the series can be identified as a nonlinear version of ARIMA(1,2,1). Neural networks trained by using second difference data are shown to give better predictions than otherwise trained by using raw data. During backpropagation training, in addition to the traditional error modification term, we also feedback the difference of two sucessive predictions in order to adjust the connection weights. Empirical results shows that the networks trained using 4-year weekly data is capable of predicting up to 6 weeks market trend with acceptable accuracy.

1. Introduction

Neural Networks, unlike traditional expert systems with explicit rules to learn knowledge, can be trained directly by feature data extracted from samples. This means that neural networks can model the behavior of known systems without being given any rule or models. Moreover, neural networks may be considered as flexible nonlinear parameterised models where the parameters may be adapted according to the available data.

In Taiwan, many people have tried to model the behavior of the stock market. Although the Random Walk Theory claims that the change of stock price are independent of its history, and we cannot obtain any indication to predict future price trends from stock price history data. The stock price change is so complex that researchers have not been able to discover a suitable model to handle behavior of stock market efficiently. In this paper, using ARIMA-based recurrent networks[1] in association with feature data pre-processing, we intend to forecast the trends of Taiwan stock market.

Autoregressive Integrated Moving Average (ARIMA) model is a linear nonstationary model [2], it uses difference operator to convert nonstationary series to stationary. Due to the availability of computer softwares of ARIMA model and high performance computers, more and more forecasting problems can now be modeled easier than before. But in modeling nonlinear series, we need to turn to other approaches. Considering neural networks capable of learning nonlinear information from appropriate trainings [3,5], we develop a ARIMA-based

0-7803-3210-5/96 $4.00©1996 IEEE

prediction system that uses the recurrent neural network [1] and propose a modified backpropagation trainig algorithm for the network. We also compare prediction performances for the ARIMA-based recurrent neural network, using raw data and second difference data respectively.

2. Feature Extraction

Many time series (e.g., stock price) behave as through they have no fixed mean. Such nonstationary behavior can be removed by performing suitable differences on the raw data of the series [2]. We define

Difference Operator ∇ : $\nabla Z_t = Z_t - Z_{t-1} = (1-B)Z_t$ $B \triangleq$ Backward shift Operator.

$$\nabla^2 Z_t = (1-B)^2 Z_t = (1-2B+B^2)Z_t$$
$$= Z_t - 2Z_{t-1} + Z_{t-2}$$

.... and so on

The autocorrelation function (ACF) is used to determine whether a series stationary or nonstationary. After certain times of differencing, one can always obtain nearly stationary series. Based on this consideration, we will apply difference operator to the series of Taiwan stock exchange weighted stock index (abbreviated as TSEWSI) for years 1991-1994. The resulting difference data are used for training the recurrent network to function as a nonlinear ARIMA(p,d,q) model. But to do this, we need to first identify the orders of p,d q.

Since we will assume all information are covered in the history, we do not use other data (such as Trading volumn, M_{IB}, Foreign Exchange, RATE, WPI). As can be seen in Fig. 1, the series is obviously nonstationary, because the ACF values do not die out after a certain length of time. We then try to second difference the series and the plot of its ACF is shown in Fig. 2, and the plot of PACF (i.e., partial ACF) is shown in Fig. 3. As indicated in Fig 2 and Fig. 3, we have identified the TSEWSI series as ARIMA(1,2,1). Note that one can also use the extended sample autocorrelation function(ESACF) [6] to identify the series.

3. ARIMA_based Recurrent Networks Forecasting System

A general linear model used for forecasting purposes is the class of ARMA(p,q)

$$x_t = r + \sum_{i=1}^{p} \varphi_i x_{t-i} \sum_{j=1}^{q} \theta_j e_{t-j} + e_t \qquad (1)$$

where it is assumed that $E(e_t | x_{t-1}, x_{t-2} \ldots\ldots) = 0$, This condition is satisfied when e_t are zero mean, independent and identically distributed, and are independent of past x_t. In this case, the minimum mean squared error predictor is $\hat{x}_t = E(x_t | x_{t-1}, x_{t-2}, \ldots, x_1)$. The optimal ARMA predictors is given by

$$\hat{x}_t = \varphi_1 x_{t-1} + \ldots.. + \varphi_p x_{t-p} + \theta_1 \hat{e}_{t-1} + \ldots.. + \theta_q \hat{e}_{t-q} \text{, where}$$
$$\hat{e}_{t-j} = x_{t-j} - \hat{x}_{t-j} \qquad j = 1,2,\ldots,q. \tag{2}$$

We develop an ARIMA-based recurrent network for prediction purpose, the stucture of the network is shown in Fig. 4. The basic architecture originates from [1], but the feature data used and the way of error feedback are different. The number of the input nodes of the recurrent network equals p+q, and number of hidden nodes is selected by a trial and error. The output of the network approximates conditional mean predictor and is given by

$$\hat{x}_t = \sum_{i=1}^{h} W_i f\left[\sum_{j=1}^{p} w_{ij} x_{t-j} + \sum_{j=1}^{q} w'_{ij}\left(\hat{x}_{t-j} - \hat{x}_{t-j-1}\right) + \theta_i\right] \tag{3}$$

where f is the sigmoidal function.

Note that in Fig. 4, the input training data (after second difference) are stationary. Instead of using $x_t - \hat{x}_t$ as in [1], we use the feedback equation $\hat{e}_t = \hat{x}_t - \hat{x}_{t-1}$ as the next new input data for an unit-delay node. The reason we do this is that any two successive data must have largest correlation, according to ACF plots shown in Fig. 2. We will use $\hat{e}_t = \hat{x}_t - \hat{x}_{t-1}$ not only during backpropagation training, but also during the recall (i.e., testing) process. We believe that the network could suffer decrease in prediction accuracy if without referring to any error data $\hat{e}_t$ during the testing phase. Furthermore, there is no way of knowing value of x_t in advance, this eliminates possibility of using $x_t - \hat{x}_t$ during the testing phase.

4. Results

In this work, ARIMA-based recurrent neural networks with 7 hidden nodes is trained by using features of TSEWSI. The series contain 1991-1994 training data and 1995 Jan-Mar testing data. To compare, we use ARIMA(1,0,1) (i.e., without differencing) and ARIMA(1,2,1) for the series. After training, we test these two networks performance for both training and testing data. The performance difference in Fig. 5 and Fig. 6 is easily seen. The predictions follow the observations nicely in ARIMA(1,2,1), but poorly in ARIMA(1,0,1). These results justify our previous arguments that the difference operations is necessary.

The predictions for 1995 Jan-Mar is shown in Fig 7. The prediction accuracy is quite good for the first six weeks. The effect of the nonlinear learning of neural networks is apparent in the error residuals of the testing set. Fig 8.a and Fig 9.a are plots of the residuals against the prediction prices for the training and testing set respectively. Fig 8.b and Fig 9.b are plots of residuals verus the previous residuals (i.e., lagging one time step) for the training and testing set respectively.

5. Conclusions

Our experimental results have shown that the ARIMA-based recurrent neural network is capable of predicting the market trend with acceptable accuracy. We also have identified the series of TSEWSI as ARIMA(1,2,1), and shown that it can be stationary after second difference operation. We believe that the prediction accuracy can be improved by adding other feature data, such as trading volume, interest rates...etc. For future works, since stock price series is a high noise dynamic system, we can try filtering [1] to cull noise or outliers, as well as studying on identification of time series.

Reference

[1] Connor, J.T.; Martin, R.D.; Atlas, L.E. "Recurrent neural networks and robust time series prediction," *IEEE Transactions on Neural Networks,* Vol. 5, Iss. 2, pp. 240-254, March 1994.

[2] George E. P. Box; Gwilym M. Jenkins; Gregory C. Reinsel , *Time Series Analysis,* Prentic Hall , 1994.

[3] James A. Freeman/David M. Skapura, *Neural Networks Algorithms, Application, and Programming Techniques,*Addison Wesley 1991.

[4] Gia-Shuh Jang; Feipei Lai; Bor-Wei Jiang; Li-Hua Chien, "The intelligent trend prediction and reversal recognition system using dual-module neural networks," *Proceedings. The First International Conference on Artificial Intelligence on Wall Street,* Vol. 26, No. 10, May. 1987.

[5] Yu, E. S.; Chen. C.Y.R., "Traffic prediction using neural networks",*IEEE Global Telecommunications Conference, including a Communications Theory Mini-Conference.* vol. 2. p.991-5, Dec 1993.

[6] Jhee, W.C,; Ro, H.B., "Decision support for ARMA model identification using hierarchically organized neural networks",*Conference Proceedings 1991 IEEE International Conference on Systems, Man, and Cybernetics.* Vol. 3, p1639-44

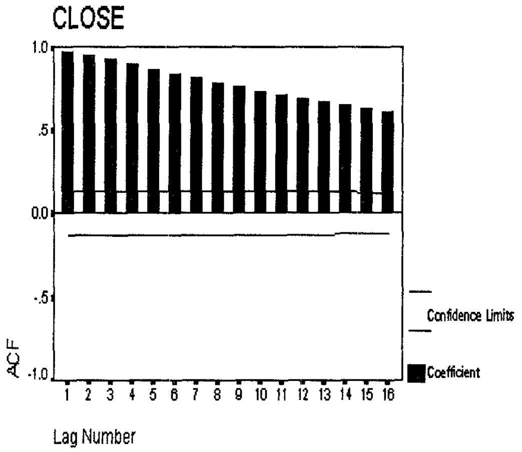

Fig. 1. ACF of TSEWSI raw data.

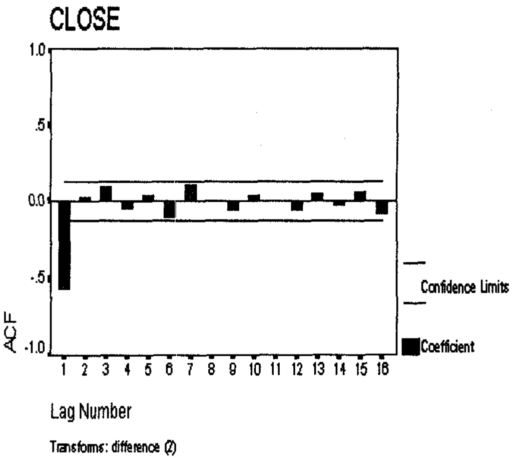

Fig. 2. ACF of TSEWSI 2nd difference data.

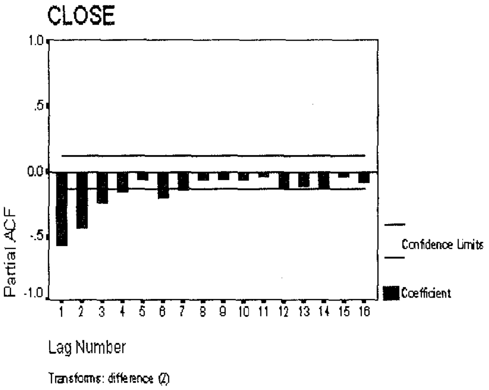

Fig. 3. PACF of TSEWSI 2nd difference data.

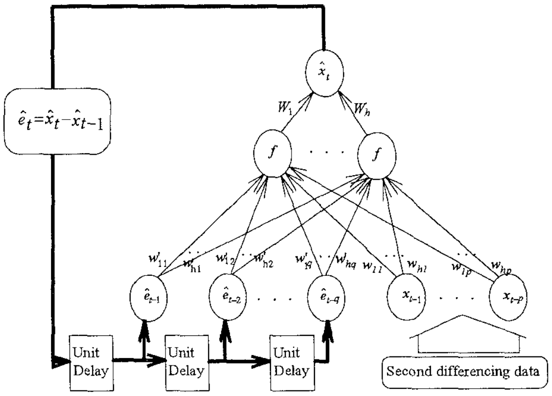

Fig. 4. The ARIMA based recurrent neural network

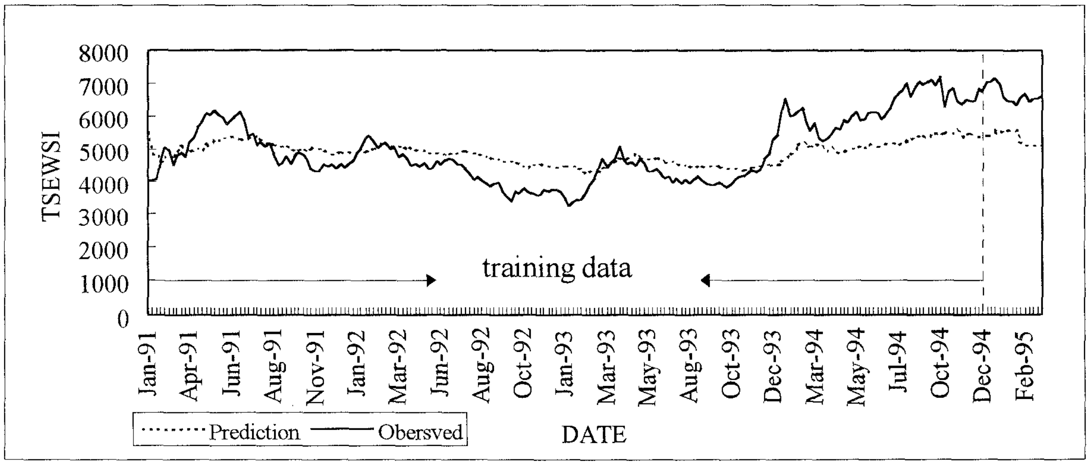

Fig. 5. Prediction performance of the network using TSEWSI raw data.

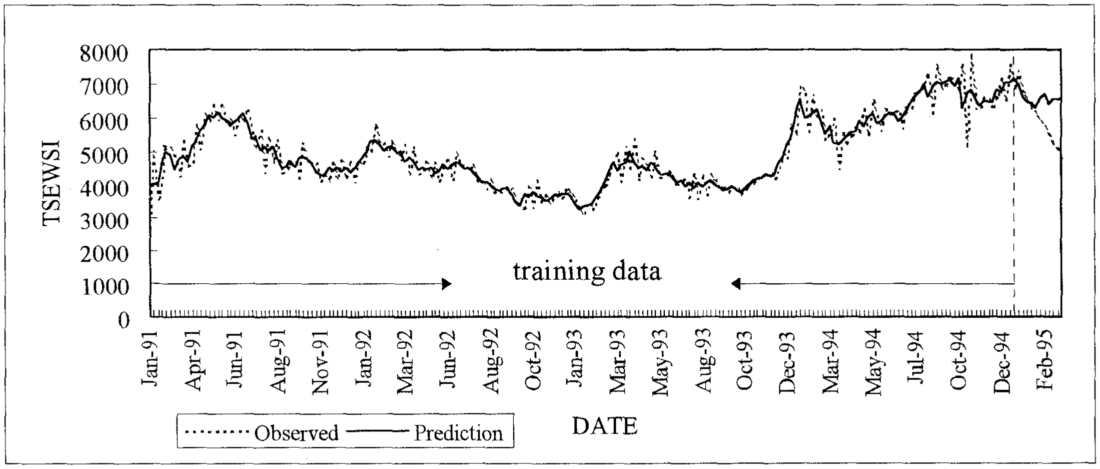

Fig. 6. Prediction performance of the network using TSEWSI 2nd difference data.

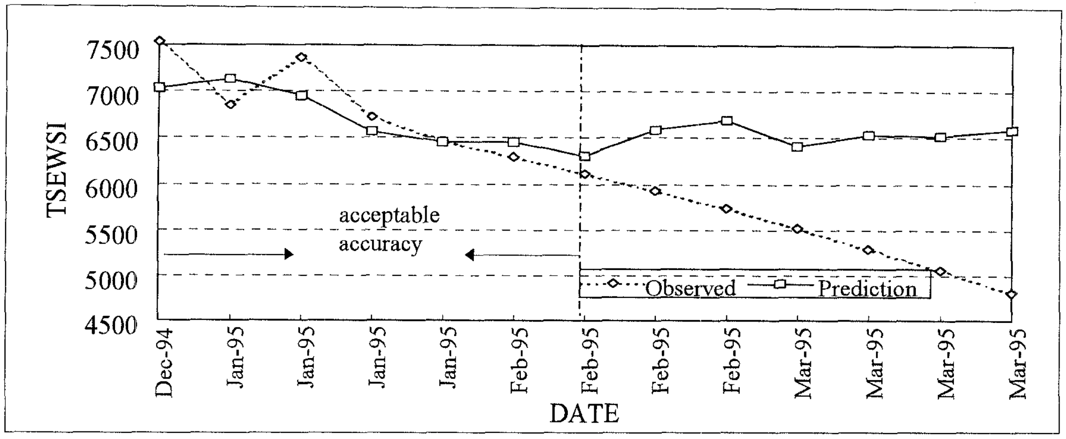

Fig. 7. ARIMA based recurrent networks using TSEWSI 2nd difference data for predicting trend during Jan /1995 -Mar/1995.

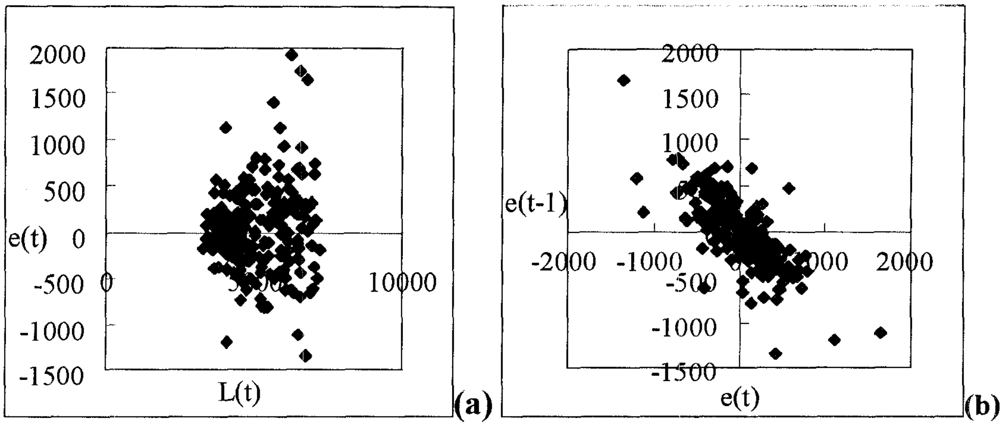

Fig 8.a Residual vs. predictions for years 1991-1994.
Fig 8.b Lag 1 scatter plot for years 1991-1994.

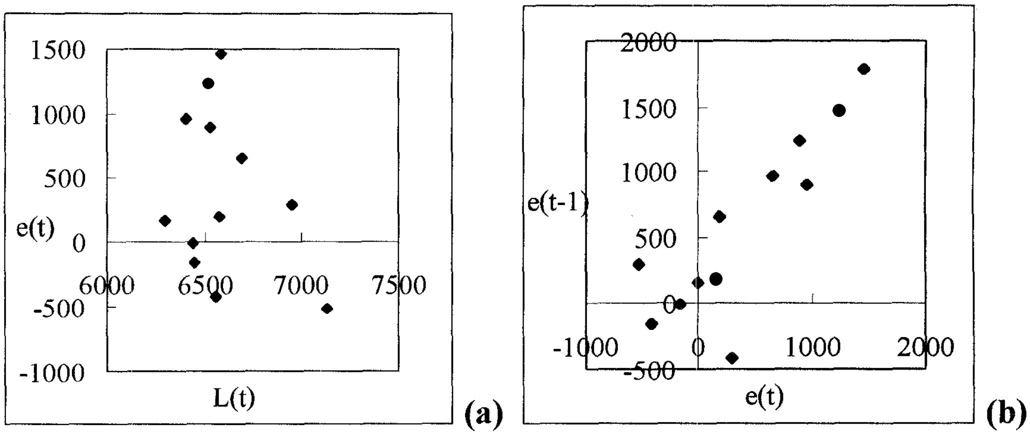

Fig 9.a Residual vs. predictions for Jan-Mar 1995.
Fig 9.b Lag 1 scatter plot for Jan-Mar 1995.