CHAPTER I

INTRODUCTION

An important application of neural networks is in time-series data prediction. Its accuracy depends upon the choice of an appropriate network and on the completeness of the collected data. In practice, those missing data occur because of malfunctioned equipment, human errors, or natural disasters. In this research, we focus on the technique of forecasting a time-series based on the data with missing values. The modeling consists of two main steps. The first step is to estimate the collected incomplete data, which are considered as *missing data* or *missing values*. The second step is to predict new data based on the statistical nature of the data obtained from the first step. The problem of estimating these missing data has recently become an extensive research topic. Managing incomplete data becomes an extensive research topic nowadays. Generally, the simple method is to ignore the missing data and to discard those incomplete cases from the data set. This approach can cause a serious problem for time-series prediction. Typically, in time-series prediction, the currently predicted values of a system depends on the historical time data of the system and can be computed by a recurrence function $x_t = f(x_{t-1}, x_{t-2}, \ldots, x_{t-k})$, where k is a constant denoting the number of previously related values. Most of the current solutions, namely maximum likelihood (ML)algorithms [1], expectation and maximization (EM) algorithms [1], [2] and multiple imputation (MI) method [3] are based on the concept of statistical probabilistic estimation. For nonlinear time-series data, k-step prediction stochastic simulation method [4]

generated the sample missing values using the distribution. For time-series analysis approach, the works by [5], [6] estimated the missing values based on ARMA models. The report of some applications [7] show the better performances of neural network models compared to the ARMA models.

For most neural approaches, the incomplete data problem is viewed as the classification problem, and is solved by supervised neural networks. Since training any supervised neural network requires both input and target data, the missing data can occur in three different aspects. The first aspect is the missing input data [8]. The second aspect is the missing target data [9]. Finally, the third aspect is both missing input and target data [10]. An example of Miller and Uyer [11] introduced an RBF classifier to estimate the missing data. Zoubin and Jordan [12] used the supervised learning and EM algorithm to improve the missing data estimation. Recently, pattern modeling and pattern characterizing have been studied in several fields such as data mining and machine learning. Clustering technique is applied for estimating missing value. Hathaway and Bezdek [13] applied the fuzzy C-means clustering to estimate the missing data of real s-dimensional data by partitioning the data sets into fuzzy clusters and estimating their cluster centers. Timm, Doring and Kruse [14] developed an extension of the Gath and Geva algorithm for assigning incomplete data points to clusters.

1.1 Motivation

The forecasting of such natural and social phenomena as hydrological cycles and climate data are very important. Data collected in practice can often be incomplete in that some data points are missing. In time-series prediction, the currently predicted value of a system depends on the historical time data of the system and can be modeled as follows: Let x_t be the value of a system at time t. The value of x_t is given by

$x_t = f(x_{t-1}, x_{t-2}, \ldots, x_{t-k})$, where k previous samples are used in the modeling. If a value x_{t-j}, $1 \leq j \leq k$, is missing, then the value of x_t may not be correctly computed by the time-series function. In this case, the estimated value of x_t, denoted by $\hat{x}_t$, must be computed from the existing x_{t-j}, $1 \leq j \leq k$. We can model $\hat{x}_t$ as $x_t + \epsilon$. The problem is to determine the technique that produces x_t such that $(\hat{x}_t - x_t)^2$ is minimized. Using a fill-in approach, those missing x_{t-j} must be estimated or filled-in first and then the functional approximation of these complete data are performed.

To achieve the best solution for this fill-in problem, several fill-in techniques must be considered. Correspondingly, an individual multilayer feedforward neural network , each of which handles one input stream, is considered. The set of complete data from each individual fill-in method is used to train a feedforward neural network to predict future values. The ensemble construction is used to combine the outputs of the individual networks to produce the best prediction output. We should show that the prediction ac- curacy can be significantly improved through ensembling a number of individual neural networks. Moreover, we propose a new fill-in technique that is improved for estimating missing values based on clustering technique for characterizing the pattern of incom- plete time-series data. The variation of time-series can be characterized as time-series pattern. The assumption is from the observation that nature phenomenon can repeat itself several times with similar characteristics. Hence, some missing data in a phenomenon can be imputed by searching and comparing with some other similar phenomena. This approach is appropriate for imputing the missing time-series data.

1.2 Objectives

In this dissertation, our objectives are as follows:

1. To propose the new fill-in technique based on the characteristics of the data;

2. To develop a neural network model for forecasting incomplete time-series data.

1.3 Area of interest

The scope of our work are as follows:

1. This research focuses on periodic time-series data.

2. A univariate time-series X is an ordered sequence of observations at equally spaced time intervals.

$$X = \{x_t, t = 1, \ldots, N\}, \tag{1.1}$$

where t is a time step, N is the number of the observed data. Some time steps of data are missing.

3. We consider both of the synthetic time-series data and various real world time-series data as follows:

3.1. Mackey-Glass chaotic time-series

This data set is available from the Internet [15] (see Figure 5.1(a)). Each data point can be mathematically generated with a constant variance (Appendix A). We can predict $x(t+6)$ from the past values of this time series, that is, $x(t-18), x(t-12), x(t-6)$, and $x(t)$. This data set is selected because this series has almost periodical behavior. The period of this series is 24 time steps. Its signal is almost stationary. This series has a constant location and variance. There are a total of 1,200 observations in our experimentation.

3.2. The annual sunspots from A.D. 1700 to A.D. 1994

This series is obtained from the Internet [16]. and represents the number of sunspots which has been recorded from the surface of the sun (see Figure 5.1(b)). The period of this series is 12 years. This univariate time-series data is selected because it is a real world case study and its signal also shows periodical behavior. Its variances are not

stable in each period. Its signal is more difficult than Mackey-Glass chaotic time series. There are a total of 1,070 observations in our experimentation.

3.3 The daily gauge height at Ban Luang gauging station, Mae Tun stream, Ping river, Thailand

This univariate time-series data is made available to us by the Royal Irrigation Department of Thailand. The gauge height data is measured at every 3 hours in a day. In our experiment, we prepared this data set to the daily gauge height data set. It is selected because it is less structured than either the Mackey-Glass chaotic time series or the monthly sunspots data (see Figure 5.1(c)). There are some high peaks at some time steps and there are some parts of data decrease gradually. There are a total of 2,000 observations in our experimentation.

3.4. The daily air temperature at Nakhon Ratchasima province, Thailand

This real-world data set is provided to us by the Meteorological Department of Thailand. It presents the most difficult problem in our case studies because of the sharp rises and falls in the series (see Figure 5.1(d)). We used a total of 2,000 observations in our experiments.

4. A feedforward neural network structure is used for prediction of time-series data.

5. A clustering technique is used for characterizing the repeated patterns of a time-series data.

1.4 Performance Measure

In our dissertation, we used two kinds of performance indexes. First, the prediction performance indexes are used for evaluating the prediction accuracy of incomplete time-series data. Second, the estimating missing value is evaluated by the imputation performance index.

1.4.1 Prediction Performance Index

The prediction performance of a network is evaluated by measuring the difference between the mean square error of the test networks and the reference network. We choose a network, which is trained by the complete training set with no missing data as the reference network. Because the occurrence of data sample dropouts is stochastic in nature, we make several runs for each fill-in method. The performance index is defined as

$$P_d = \frac{1}{R}\left(\sum_{i=1}^{R} E_i^T - \sum_{j=1}^{R} E_j^{RN}\right) \tag{1.2}$$

where E_i^T denotes the mean square error of the tested network, E_j^{RN} denotes the mean square error of the reference network, and R denotes the number of runs per fill-in method. The parameter R is twenty-five in our experiments. The interpretation of P_d is that if the prediction performance of the reference network is worse than that of the test network, then P_d is less than zero, and the lower value shows the better performance. If the prediction performance of the reference network is better than that of the test network, then P_d is greater than zero. Otherwise P_d is zero.

In case we focus on the average performance index, then, the performance index is defined as

$$P_d' = \frac{1}{L \times R}\left(\sum_{i=1}^{L\times R} E_i^T - \sum_{j=1}^{L\times R} E_j^{RF}\right) \tag{1.3}$$

where L denotes the number of different percentage of missing data, and E_j^{RF} denotes the mean square error of the proposed network. The meaning of P_d' is that if the average prediction performance of the proposed network is worse than that of an individual network, then P_d' is less than zero. If the average of prediction performance of proposed network is better than that of the individual network, then P_d' is greater than zero. Otherwise, P_d' is zero.

1.4.2 Imputation Performance Index

We measure the error of estimating missing values between the actual values x and the prediction values $\widehat{x}$ by using the mean squared error (MSE)

$$\text{MSE} = \frac{1}{R \times M} \sum_{i=1}^{R} \sum_{j=1}^{M} (x_{ij} - \widehat{x}_{ij})^2 \tag{1.4}$$

where M is the the number of missing value, and R denotes the number of runs per fill-in method. We set R to 30 in our experiments to average out variations due to the stochastic nature of choosing which values to be considered missing. The lower MSE shows the better prediction of missing value.

Another performance measure is Pearson's correlation (CORR) between the actual values and the estimated values:

$$\text{CORR} = \frac{\sum_{i=1}^{R} \sum_{j=1}^{M} (N_{ij}^{Actual} \times N_{ij}^{Predict})}{\sqrt{\sum_{i=1}^{R} \sum_{j=1}^{M} (N_{ij}^{Actual} \times N_{ij}^{Actual})} \sqrt{\sum_{i=1}^{R} \sum_{j=1}^{M} (N_{ij}^{Predict} \times N_{ij}^{Predict})}}. \tag{1.5}$$

where

$$N_{ij}^{Actual} = x_{ij} - \frac{1}{R \times M} \sum_{i=1}^{R} \sum_{j=1}^{M} x_{ij}, \tag{1.6}$$

and

$$N_{ij}^{Predict} = \widehat{x}_{ij} - \frac{1}{R \times M} \sum_{i=1}^{R} \sum_{j=1}^{M} \widehat{x}_{ij}, \tag{1.7}$$

are the normalized zero mean of the actual values and the estimated values. The coefficient CORR measures the degree of similarity between the actual values and the estimated values.

The average imputation performance index P_{Imp} by the varied window clustering (WDC) algorithm are evaluated by measuring the difference between the MSE of the WDC algorithm and that of a reference method to which we are comparing the WDC. This performance index is adopted from our previous work [17] and is defined as

$$P_{Imp} = \frac{1}{L} \left(\sum_{i=1}^{L} E_i^{RM} - \sum_{j=1}^{L} E_j^{WDC} \right) \tag{1.8}$$

where L denotes the number of different percentage of missing data, E_j^{WDC} denotes the mean square error of the WDC algorithm, E_i^{RM} denotes the mean square error of the reference method. The interpretation of P_{Imp} is such that if the imputation performance index of the WDC algorithm is worse than that of the reference method, then P_{Imp} is less than zero. If the average imputation performance of the WDC algorithm is better than that of the reference method, then P_{Imp} is greater than zero. Otherwise P_{Imp} is zero.

1.5 Contributions of the Dissertation

In this dissertation we instantiate the solutions of the above objectives in a novel method for a fill-in technique and develop the new model of the neural network of incomplete time-series prediction.

The proposed fill-in technique has several interesting features:

- this concept is easy to implement;
- we need not to require the distribution of time-series data; and
- it can be used with non-stationary time-series data.

And we proposed a new model of the neural network for forecasting incomplete time-series data. Furthermore, the proposed neural network model is a solution for improving the prediction accuracy.

1.6 Dissertation Organization

The rest of the dissertation is organized into four additional chapters. In Chapter 2, we review the managing incomplete data and time-series prediction. In Chapter 3, we present both of FI-GEM network and RMD-FSE network for incomplete time-series

prediction. In Chapter 4, we present a new imputation of missing value, namely varied window clustering (WDC) algorithm and the experimental results. In Chapter 5, we discuss and conclude the dissertation.