CHAPTER II

INCOMPLETE TIME-SERIES PREDICTION

2.1 Managing Incomplete Data

There are three simple methods for handling missing data. The first method is to ignore the missing data, and to discard those incomplete cases from the data set. The second method is called imputation, which attempts to fill in with some plausible values. The third method uses the interpolation technique. The examples of interpolation technique have been used by Hughes and Smakhtin [18] for estimating missing observations in hydrological data. However, those methods are not generally good and can cause a serious problem of prediction accuracy. The analysis of missing values based on statistical approach have been studied since the early 1970s. Little and Rubin [2] provided a coherent theory based on likelihoods derived from statistical models for the missing data problem. The famous expectation and maximization (EM algorithm), which is a general technique that finds maximum likelihood estimates for incomplete data, were proposed by Demspster, Laid and Rubin [1] in 1977. The extension of EM as the regularized EM algorithm [19] was proposed for climate data. For the other approach, Hathaway and Bezdek [13] applied the fuzzy c-means clustering to estimate the missing data.

In this work, we consider eight techniques for filling in the missing data, which are (1) cubic smoothing spline interpolation, (2) multiple imputation by the EM algorithm: random selection, (3) multiple imputation by the EM algorithm: average selection, (4) The regularized EM algorithm: random selection, (5) the regularized EM algorithm:

average selection, (6) k-segment principal curves [20], (7) the Optimal completion strategy fuzzy c-means clustering, and (8) WDC clustering. (1)-(6) are applied to the same data set with some missing data in [17], [21], and [22]. (7) is used for performance evaluation with our proposed technique (8) in [23]. Prior to the filling processes, the data preparation and interpretation are established for each technique as follows.

2.1.1 Cubic Smoothing Spline Interpolation

The data set is organized as a univariate data with equally space on x-axis. If some value x_t at time t is missing, then the value of t on x-axis is also skipped. The accuracy of the spline interpolation is measured by the following cost function.

$$P\sum_t w_t(x_t - f(t))^2 + (1-P)\int (D^2 f)^2 \tag{2.1}$$

where P is a constant, $0 \leq P \leq 1$, w_t is the weighting factor of datum x_t, $f(t)$ is a cubic spline, and $D = I_{n \times n}$ is an identity matrix of size $n \times n$. In our experiment we set $w_i = 1$ and $P = 0.99$. The approximated value of the missing value x_t is given by

$$\hat{x_t} = f(t) \tag{2.2}$$

2.1.2 Multiple Imputation by EM Algorithm: Random Selection

This EM algorithm is based on the concept of Schafer [3]. The inputs to EM algorithm are partitioned and arranged in a time-series form with a window size of k. Each set of inputs is, then, stacked to form an input matrix A as shown in equation (2.3). Element x_t denotes an input value at time t. The value of each missing datum is set to some special values, such as -9999. After the multiple imputation process, all missing data are estimated and a reconstructed matrix $A^{EM} = (a_{i,j})$ is produced.

$$A^{EM} = \begin{bmatrix} x_1 & x_2 & \cdots & x_k \\ x_2 & x_3 & \cdots & x_{k+1} \\ x_3 & x_4 & \cdots & x_{k+2} \\ \cdots & \cdots & \cdots & \cdots \\ x_{N-k+1} & x_{N-k+2} & \cdots & x_N \end{bmatrix} \tag{2.3}$$

Suppose x_t is a missing datum. Obviously, the number of x_t can appear diagonally from 1 to k times. Let the number of appearances of x_t in A be k'. After EM algorithm, there are k' possible estimated values of $\hat{x}_t$. Hence, we randomly select one of the values of $\hat{x}_t$ as the fill-in value.

2.1.3 Multiple Imputation by EM Algorithm: Average Selection

The process is similar to that of EM algorithm in Section 2.1.2. The difference is the estimated value of $\hat{x}_t$ is computed by averaging all its possible values.

2.1.4 Imputation by the regularized EM algorithm: Random Selection

In reference [19], the author proposed the method to estimate the mean and the covariance matrix of an incomplete data set and fill-in missing values with imputed values by the regularized EM. The regularized EM algorithm is based on iterated analysis of linear regressions of variables with missing values with regression coefficients estimated by ridge regression, a regularized regression method in which a continuous regularization parameter controls the filtering of the noise in the data [19]. After imputation process, a reconstructed matrix is produced. Similar to Section 2.1.2, the estimated value of $\hat{x}_t$ is randomly selected from all of its estimated values.

2.1.5 Imputation by the regularized EM algorithm: Average Selection

The estimation process is the same as that in Section 2.1.4 except that the selection of the value of $\hat{x}_t$ is performed by averaging all of its estimated values, similar to the selection process in Section 2.1.3.

2.1.6 k-Segment Principal Curves

The k-segment principal curve interpolation method is an incremental one that finds principal curves which are generalizations of principal components [20]. Line segments are inserted and fitted to form polygonal lines on the univariate time-series data. The missing value is obtained by evaluating the interpolation on those combined polygonal segments of time-series data.

2.1.7 Fuzzy C-means Clustering

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 of their cluster centers. Four strategies are proposed for doing FCM clustering of incomplete data sets. We selected a strategy, which is optimal completion fuzzy C-means (OCSFCM) for testing with our four case studies. FCM approach is compared with our proposed by measuring the accuracy of estimating incomplete data.

2.2 Time-Series Prediction

2.2.1 Multilayer Feedforward Neural Network

Neural networks are emerging as new models for prediction of nonlinear phenomena in 1980. Neural networks are powerful when applied to highly-complex problems. A multilayer feedforward neural network (MLP) is a powerful structure for prediction.

The nonlinearity is distributed through certain layers of processing. In Figure 2.1, a multilayer feedforward neural network is shown. The input samples are fed to the input layer. The prediction is produced at the output layer. The output of each layer are connected to the adjacent layer.

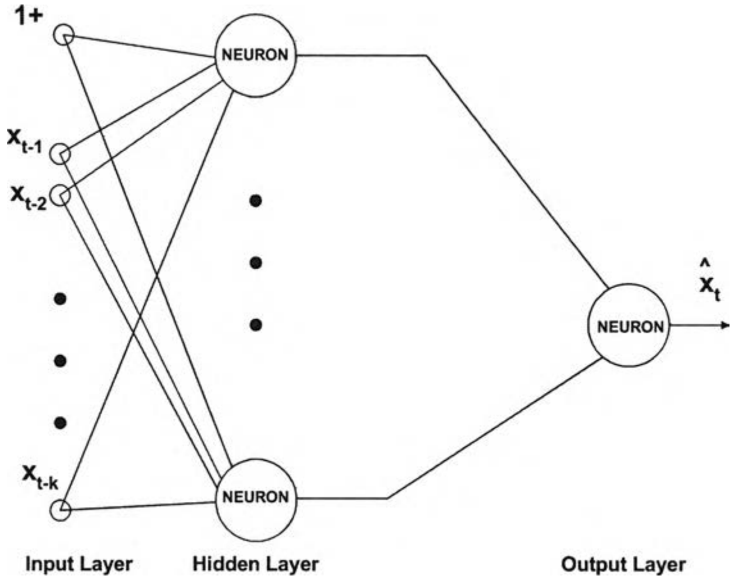

Figure 2.1: Multilayer Feedforward Neural Networks

On the problem of time-series prediction, the inputs must capture the time evolution of the underlying discrete time random signals. An input is time-delayed, i.e. $x(i), i = 0, 1, 2, \ldots, k)$. The overall predictor can be represented as

$$x_t = f(x_{t-1}, x_{t-2}, \ldots, x_{t-k}), \tag{2.4}$$

where f represents the nonlinear mapping of the neural network, k is the previous samples, which are used in the modeling, and t is time step.

2.2.2 Ensemble Neural Network Design

Even though a single neural network can be efficiently used for the prediction of time-series data, a combination of many neural networks of the same type significantly shows the improvement of the prediction performance. Ensemble networks consist of independently trained neural networks which are combined as a single master network. The network is used as the second estimating step. The input to each sub-network is the output from each missing data fill-in technique. The Generalized Ensemble Method (GEM) proposed by Perrone [24] is a general technique for combining the outputs of all individual neural networks. An equation of this method is shown as follows:

$$f_{GEM} = \sum_{i=1}^{N^{GEM}} \alpha_i f_i(x), \tag{2.5}$$

where N^{GEM} is the number of individual networks, $f_i(\mathbf{x})$ is the output value of network i, and α_i is the weighting parameter for network i. All α_is must satisfy the constraint of $\sum \alpha_i = 1$. Each α_i is defined by

$$\alpha_i = \frac{\sum_j C_{ij}^{-1}}{\sum_k \sum_j C_{kj}^{-1}} \tag{2.6}$$

where C_{ij} are the elements of the covariance matrix of the errors from the function estimators f_i and f_j. C_{kj} are the elements of the covariance matrix of the errors from the function estimators f_k and f_j.