View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Dogus University Institutional Repository

Journal of Applied Sciences 7 (6): 922-925, 2007
ISSN 1812-5654
© 2007 Asian Network for Scientific Information

Reconstruction of Time Series Data with Missing Values

Mitat Uysal
Dogus University, Acibadem-Kadikoy 34722, Istanbul, Turkey

Abstract: Missing data are a part of almost all research and it must be decided how to deal with it from time to time. Missing data creates several problems in many applications which depend on good access to accurated data. Conventional methods for missing data, like listwise deletion or regression imputation, are prone to three serious problems: Inefficient use of the available information, leading to low power and Type II errors. Biased estimates of standard errors, leading to incorrect p-values. Biased parameter estimates, due to failure to adjust for selectivity in missing data. In this study, we propose a new algorithm to predict missing values of a given time series using Radial Basis Functions.

Key words: Handling missing data, time series, radial basis functions, function approximation, forecasting modeling, simulation

INTRODUCTION

Time series data are used to represent many real world phenomenon. For various reasons, a time series database may have some missing data. Traditional interpolation or estimation methods usually become invalid when the observation interval of the missing data is not small (Hong and Chen, 2003).

The methods of handling missing data are directly related to the mechanisms that caused the incompleteness. These mechanisms fall into three classes (Sentas and Angelis, 2005; Little and Rubin, 2002).

- Missing Completely at Random (MCAR): The missing values in a variable are unrelated to the values of any other variables, whether missing or valid.
- Non-Ignorable Missingness (NIM): The probability of having missing values in a variable depends on the variable itself.
- Missing at Random (MAR): This can be considered as an intermediate situation between MCAR and NIM. The probability of having missing values, does not depend on the variable itself but on the values of some other variable.

Missing data techniques are given in Little and Rubin (2002). They can be listed as: Listwise deletion, mean imputation, regression imputation and expectation maximization. Details can be obtained from Little and Rubin (2002).

Many recent publications appeared in literature related to dealing missing data.

Choi and Kim (2002) presented a physics-based approach for automatically reconstructing three dimensional shapes in a robust and proper manner from partially missing data.

Tang and Hung (2006) have proposed an algoritbm to estimate projective shape, projective depths and missing data iteratively.

Yemez and Wetherilt (2007) presented a hybrid surface reconstruction method that fuses geometrical information acquired from silhouette images and optical triangulation.

Golyandina and Osipov (2007) have proposed a method of filling in the missing data and applied to time series of finite rank.

Heintzmann (2007) introduced a novel way of measuring the regain of out-of-band information during maximum likelihood deconvolution and applied to various situations.

Formal representation of missing data: Original data matrix $D = (d_{ij}) I = 1,2,3\ldots n, j = 1,2,\ldots k$ contains time series data where d_j is the value of variable d_j for case I.

When there are missing data, the missing data indicator matrix $M = (m_{ij})$ can be defined as below:

if $m_{ij} = 1$ then d_j is missing
if $m_{ij} = 0$ then d_{ij} is present
(Sentas and Angelis, 2005).

Radial basis functions for time series forecasting: An RBF network consists of 3 layers: an input layer, a hidden layer and an output layer. A typical RBF network is shown in Fig. 1.

Mathematically, the network output for linear output nodes can be expressed as below:

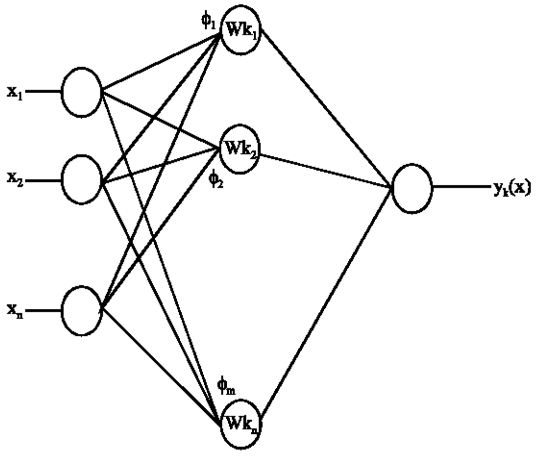


Fig. 1: Typical RBF network

$$y_k(x)=\sum_{j=1}^{m} w_{kj}\Phi_j\left(\left\|x-\vec{x}_j\right\|\right)+w_{k0}$$

Where x is the input vector with elements x_i (where I is the dimension of the input vector),

$\vec{x}_j$ is the vector to determine the center of the basis function Φ_j with elements $\vec{x}_{ji}$, w_{kj}'s are the weights and w_{k0} is the bias (Harpham and Dawson, 2006). The basis function Φ_j (-) provides the nonlinearity. The most used basis functions are Gaussian and multiquadratic functions (Harpham and Dawson, 2006).

Calculating the optimal values of weights: A very important property of the RBF Network is that it is a linearly weigthed network in the sense that the output is a linear combination of m radial basis functions, written as below:

$$f(x)=\sum_{i=1}^{m} w^{(i)}\Phi^{(i)}(x)$$

(Duy and Chong, 2003)

The main problem is to find the unknown weights $\{w^{(i)}\}_{i=1,m}$ For this purpose, the general least squares principle can be used to minimize the sum squared error:

$$SSE=\sum_{i=1}^{n}\left[y^{(i)}-f(x^{(i)})\right]^2$$

With respect to the weights of f, resulting in a set of m simultaneous linear algebraic equations in the m unknown weights

$$(B^T B)w=B^T y$$

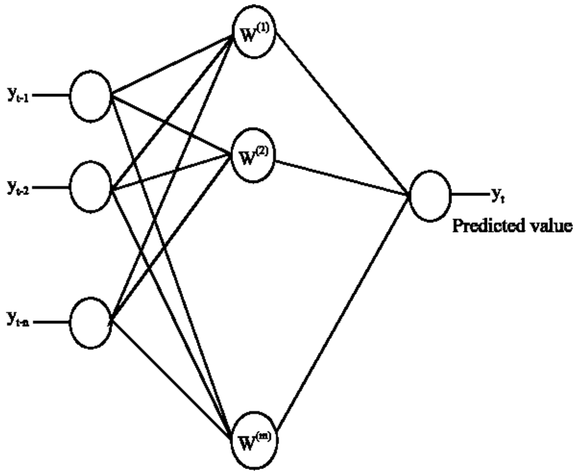


Fig. 2: Finding the predicted value y_t

where

$$B=\begin{bmatrix}\Phi^{(1)}(x^{(1)}) & \Phi^{(2)}(x^{(1)}) & . & \Phi^{(m)}(x^{(1)})\\ \Phi^{(1)}(x^{(2)}) & \Phi^{(2)}(x^{(2)}) & . & \Phi^{(m)}(x^{(2)})\\ . & . & . & .\\ \Phi^{(1)}(x^{(n)}) & \Phi^{(2)}(x^{(n)}) & . & \Phi^{(m)}(x^{(n)})\end{bmatrix}$$

$$w=[\,w^{(1)}, w^{(2)}, \ldots\ldots w^{(m)}]^T,\ y=[\,y^{(1)}, y^{(2)}, \ldots\ldots y^{(n)}]^T$$

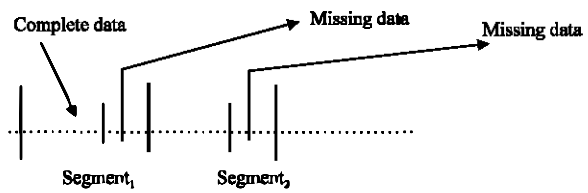
In the special case where n = m the resultant system is just

$$Bw=y \qquad \text{(Duy and Chong, 2003)}$$

The output y(x) represents the next value of y in time t taking input values x_1, x_2, ...x_n that represent the previous function values set of the time series with values y_{t-1}, y_{t-2}, ...y_{t-n}. So, x_n corresponds to y_{t-1}, x_{n-1} corresponds to y_{t-2} etc. as in Fig. 2.

Reconstruction of data series by radial basis functions: a new algorithm: The following algorithm is proposed in this work to find the values of missing data.

- Remove the 20% of the original data from the data set. Divide the data set into segments so that each segment contains some missing data:



- Use the complete data of segment$_i$ to find an artificial time series equation with an RBF network that means finding the weights in the RBF approximation.
- Calculate the error in each segment according to the following formula:

$$e_i^j = y(x_i)^j - r_i^j$$

Where e_i^j is the error value in the x_i point on the jth segment.

- Calculate the sum squared errors in each segment in each pass of the algorithm.

$$SSE_k = \sum_{j=1}^{n}\sum_{i=1}^{m}(e_i^{\,j})^2$$

where k is the number of the pass.

- Replace the missing data with the predicted values in each segment in the pass m where SEE_m is the minimum value of SSE_k. Stop the algorithm.

SIMULATION RESULTS

Several simulation runs were carried out in a computer environment to find the optimal values of parameters in radial basis functions like width δ and centers $(\bar{x}_i\text{'s})$ to obtain good predictions for the missing data in the time series.

Figure 3 shows the results of the first simulation run. In this run, the first 40 data items were used to predict the next 8 data items that was considered missing data and the results were compared with the real data. Real data values are represented with symbol + and predicted values are represented with symbol o.

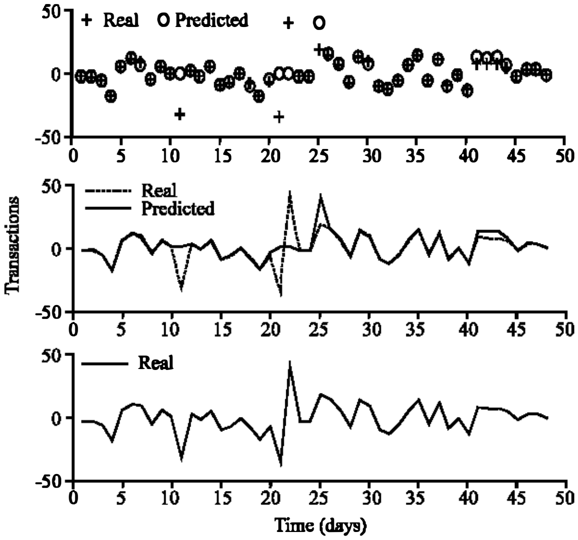


Fig. 3: Gaussian Function sigma = 0.93 and 18 neurons in the hidden layer

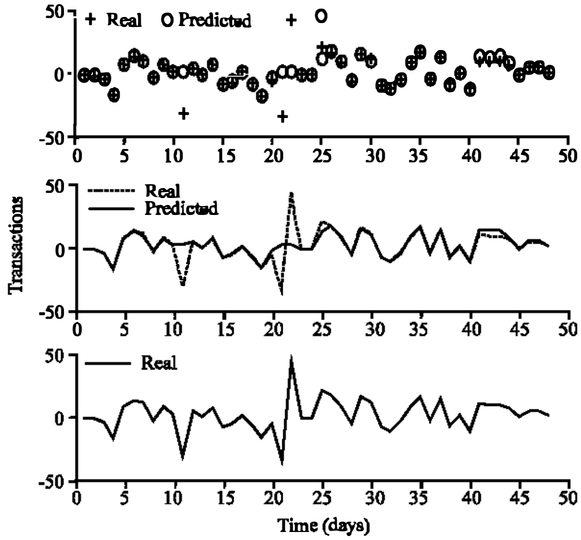


Fig. 4: Gaussian Function sigma = 1 and 18 neurons in the hidden layer

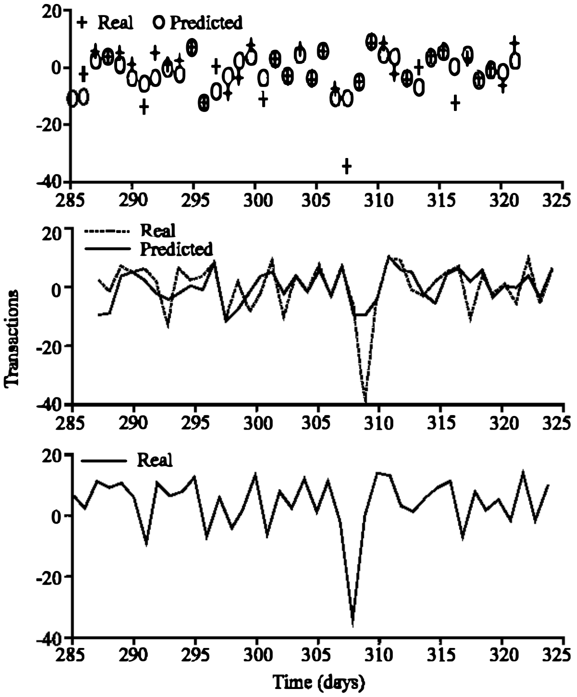


Fig. 5: Gaussian Function sigma =1 and 18 neurons in the hidden layer for the last 40 data

In Fig. 4, similar experiment was carried out with δ = 1 for a Gaussian function and better results were obtained.

Figure 5 shows, the results of the similar experiment for the last 40 data items for a Gaussian function.

CONCLUSIONS

In this study, I proposed a new algorithm to predict missing values of a given time series using Radial Basis Functions. Radial Basis Functions provide a good way to predict the values of missing data in a time series. In this study, a monthly data log of a bank was used to carry out the simulation experiments. The data log file consisted of 324 data items. This file was divided to small parts with 48 data items for the first 6 parts and 36 data items for the last part. The last 20% of the data for each part was removed and these removed data items were predicted using RBF's and the 80% of the data items for each part. For some optimal parameters of the RBF's, very good predictions are obtained for the missing data.

REFERENCES

Choi, S.M. and M.H. Kim, 2002. Shape reconstruction from partially missing data in modal space. Comput. Graphics, 26: 701-708.

Duy, N.M. and T.T. Cong, 2003. Approximation of function and its derivations using Radial Basis Function Networks. Applied Math. Modelling, 27: 197-220.

Golyandina, N. and E. Osipov, 2007. The Caterpillar-SSA method for analysis of time series with missing values. J. Stat. Plan. Inference, (In Press).

Harpham, C. and C.W. Dawson, 2006. The effect of different basis functions on a radial basis function network for time series prediction: A comparative study. Neurocomputing, 69: 2161-2170.

Heintzmann, R., 2007. Estimating missing information by maximum likelihood deconvolution. Micron, 38: 136-144.

Hong, B. and C.H. Chen, 2003. Radial basis function neural network-based nonparametric estimation approach for missing data reconstruction of non-stationary series. IEEE Int. Conf. Neural Networks and Signal Processing Nanjing, China, December 14-17, pp: 75-78.

Little, R.J.A. and D.B. Rubin, 2002. Statistical Analysis with Missing Data. John Wiley Publishers Company.

Sentas, P. and L. Angelis, 2006. Categorical Missing data imputation for software cost estimation by multinomial logistic regression. J. Syst. Software, 79: 404-414.

Tang, W.K. and Y.S. Hung, 2006. A subspace method for projective reconstruction from multiple images with missing data. Image Vision Comput., 24: 514-525.

Yemez, Y. and C.J. Wetherilt, 2007. A volumetric fusion technique for surface reconstruction from silhouettes and range data. Comput. Vision Image Understanding, 105: 30-41.