Adaptive Estimation of Missing Environmental Parameters Based on Radial Basis Function Neural Networks

Anuj Kumar and Hiesik Kim

***Abstract*—This paper has been presented the adaptive estimation for missing environmental parameters for short duration. The Radial Basis Function based Artificial Neural Network technique has been discussed and used this technique the estimation of the missing environmental parameters. This work assumes that data are missing completely at random. This implies that we expect the missing values or input vector to be deducible in some complex manner from the remaining data. Two cases of missing parameters have been considered, in first case one parameter is missing, and in second case two parameters are missing.**

***Index Terms*—RBFNN, artificial neural network, missing data, random data.**

I. Introduction

Real time processing applications are highly dependent on data acquisition and therefore quite often suffer from the problem of missing input variables. Databases such as those which store measurement or environmental data may be subjected to missing value or variable either in data acquisition or data storage process [1]. There are several reasons, why the data may be missing. They may be missing because one or more than one sensor may have temporarily malfunctioned, or the data may not have been entered correctly or a break in the data transmission line [2]. Missing data has difficulty in the analysis and decision making processes which depend on these data; no matter how accurate and efficient are the methods of estimation. To overcome this issue, various techniques have been proposed to find the missing data, can be reported as [3]-[11].

However, some statistical methods, like mean substitution, and hot deck imputation have a high likelihood of producing biased estimates or make assumptions about the data that may not be true, affecting the quality of decisions made based on the data.

To predict the exact values of the missing variables, a proper estimation method needs to be selected.

In this paper, the RBF based ANN technique has been proposed as a solution to the problem of missing data for short duration. Radial basis function (exact fit) approach has been used for ANN training and test. Further the radial basis neural networks can be designed directly by fitting special response inputs where they will do the most good. The applicability of the Graphical User Interface (GUI) of Neural Network tool box under MATLAB environment has been explored.

II. Artificial Neural Network Techniques

A. Basics of Artificial Neural Networks (ANN) and Radial Basis Function (RBF)

An ANN is a computational model of the brain. The ANN$_S$ assume that computation is distributed over several simple units called neurons, which are interconnected and operate in parallel, thus known as parallel distributed processing systems or connectionist systems. Implicit knowledge is built into the ANN by training it. The ANN captures the domain knowledge from the examples. ANN can handle continuous as well as discrete data and has good generalization capability. Several types of ANN structures and training algorithms have been proposed in [12], [13]. The transfer function for a radial basis neuron is:

$$radbas(n) = e^{n^2} \tag{1}$$

Here *n* is the net input to the *radbas* transfer function and it is defined as the vector distance between its weight vector and the input vector, multiplied by the bias. This *radbas* function calculates a layer's output from its net input.

B. Radial Basis Function (RBF) Based ANN

In radial basis function (RBF) based ANN; the learning is equivalent to finding a surface in a multi dimensional space that provides a best fit to the training data, with the criterion for best fit being measured in some statistical sense. Radial basis networks may require more neurons than standard feed-forward back propagation networks, but often they can be designed in a fraction of the time it takes to train standard feed-forward networks. They work best when many training vectors are available. The basic form of RBF architecture involves entirely three different layers. The input layers is made up of source nodes while the second layer is hidden layer of high enough dimension which serves a different purpose from that in a multilayer perceptron. Finally the output layer supplies the response of the network to the activation patterns applied to the input layer. The transformation from the input layer to hidden is nonlinear whereas the transformation from the hidden unit to the output layer is linear [13]-[16].

Manuscript received September 12, 2012; revised November 33, 2012.

The authors are with the Electrical and Computers Engineering Department, University of Seoul, South Korea (e-mail: anujkumariitd@gmail.com, drhskim@uos.ac.kr).

DOI: 10.7763/IJCTE.2013.V5.685

C. MATLAB Based Neural Network Toolbox Graphical User Interface

Neural network toolbox provides tools for designing, implementing, visualizing, and simulating neural networks. Neural networks are invaluable for applications where formal analysis would be difficult or impossible, such as pattern recognition, and nonlinear system identification and control. Neural network toolbox software provides comprehensive support for many proven network paradigms, as well as graphical user interface that enable to design and manage given networks. The modular, open, and extensible design of the toolbox simplifies the creation of customized functions and networks.

The graphical user interface is designed to be simple and user friendly, but we will go through a simple example to get started.

The GUI allows creating networks, entering data into the GUI, initialize, train, and simulating networks, exporting the training results from the GUI to the command line workspace, import data from the command line workspace to the GUI, for the opening of the Network/Data Manager Window, the command is 'nntool'.

III. Methodology

The selections of the method are dependent on the nature of the missing data and the accuracy required.

A. The Nature of Missing Data

If there are data sets with variables $X = \{X_1, X_2, \ldots, X_N\}$, where $X_1, \ldots, X_N$ are some input variables and if X_1 or X_2 or X_N or X_1X_2 or $X_1X_2, \ldots, _N$ are missing input variables. The nature of the missing data can be characterized in three categories as follows [17]-[19]:

1) Missing completely at random (MCAR)

There are several reasons why the data may be missing. They may be missing because one or more than one sensor malfunctioned, the data were not entered correctly, a break in data transmission line, etc. Here the data are missing completely at random (MCAR). When we say that data are missing completely at random, we mean that the probability that an observation ($X_{1\text{or}\ldots\ldots,\text{ or }N}$) is missing is unrelated to the value of any other variables.

2) Missing at random (MAR)

This occurs if the missing value for the input vector depends on other variables in the dataset, such that the pattern in which the data becomes missing is traceable. That is the probability of the missing data is dependent only on any input vector, the existing values in the database and not on any missing data.

3) Missing not at random (MNAR)

This occurs when the missing value for the input vector depends on the other missing values, such that the existing data in the database can not be used to approximate the missing values. This is also known as the non-ignorable case. The probability that $X_{1\text{ or}\ldots\text{ or }N}$ is missing is dependent on the missing data.

This work assumes that data are missing completely at random (MCAR). This implies that we expect the missing values or input vector to be deducible in some complex manner from the remaining data.

Initially the real time data of seven parameters for 30 days is collected and the collected data are applied to the above mentioned methods.

B. RBF Based ANN Architecture for Missing One Parameter

The inputs and outputs of ANN, structure of the its network using appropriate data should be done with utmost care for effective incipient missing parameter of the EM system. The inputs are real time measuring environmental parameters by the EM system and there is one output of the missing parameter. Therefore, as shown in Figure 1 there is six input neurons and one output neuron. For simulation, a case of a real time measured six parameters and one missing parameter is taken up by the EM system. The instantaneous values are used in the training and testing/validation process. A total of 1,16,387 data sets are used in the training. Therefore as mentioned earlier there are six input neurons and one output neurons in the proposed scheme. The architecture of the network is shown in Fig. 2.

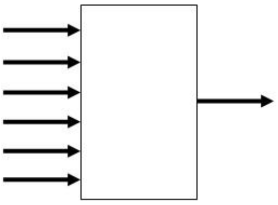

Fig. 1. Illustrates the six inputs and one output of the ANN.

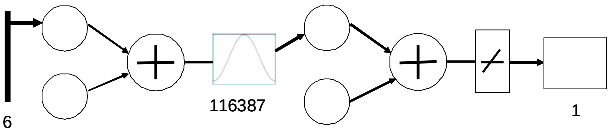

Fig. 2. Architecture of the one output network.

1) Simulation results/performance of proposed RBF ANN

The network is trained using the radial basis function (exact fit) ANN under MATLAB Neural Network tool box (GUI). The spread constant is taken as 1.0 in the present work. The trained network is tested with data sets consisting of trained data of 15 data sets (24 hours averaging). The tested results are shown in Table I. These test data sets show the expected output. It is clear from the Table I that the ANN has successfully identified the missing parameter. The output of the ANN for a particular missing parameter is exactly the same as expected.

C. RBF Based ANN Architecture for Missing Two Parameters

Here again the inputs and outputs of ANN, structure of its network using appropriate data should be done with utmost care for effective incipient missing parameter of the EM system. The inputs are real time measuring environmental parameters by the EM system and there are two outputs of the missing parameters. Therefore, as shown in Fig. 3 there is five input neurons and two output neurons.

TABLE I: RESULTS OF IMPLEMENTED METHOD FOR MISSING ONE PARAMETER

Input Parameters (24 Hours Averaging)						**Simulated Results**
CO (ppm)	CO_2 (ppm)	SO_2 (ppm)	NO_2 (ppm)	O_2 (%)	RH (%)	Temp. (^oC)
8.15875	421.6541	0.130	0.025	20.45000	69.40	31.97
8.45250	421.7654	0.130	0.020	19.29375	71.00	32.34
8.46125	427.2375	0.110	0.150	19.02250	77.00	30.17
8.39125	423.8738	0.105	0.026	19.32000	79.90	29.46
8.25125	424.3525	0.110	0.025	20.16125	81.90	28.87
8.44375	423.3925	0.116	0.020	19.22875	81.10	29.35
8.13750	423.3938	0.115	0.025	19.39500	82.40	28.96
8.09375	423.3925	0.130	0.029	19.25875	85.20	28.44
8.06750	428.6788	0.120	0.025	19.05500	84.40	29.09
8.39125	423.8738	0.105	0.026	18.64500	81.10	29.77
8.10250	428.6788	0.110	0.025	19.05500	81.40	30.05
8.10250	428.6788	0.110	0.025	19.05500	84.30	28.83
8.10250	428.6788	0.120	0.025	20.01000	85.90	28.21
8.24250	425.7963	0.100	0.021	19.25000	86.70	27.68
8.46125	427.2375	0.110	0.140	19.00200	89.10	27.40

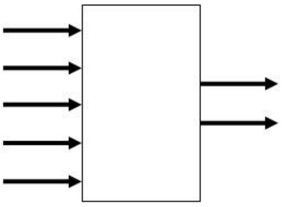

Fig. 3. Illustrates the five inputs and two outputs of the ANN.

For simulation, a case of a real time measured five parameters with two missing parameters are taken up by the EM system. The instantaneous values are used in the training and testing/validation process. A total of 1, 16, 387 data sets are used in the training. Therefore as mentioned earlier there are five input neurons and two output neurons in the proposed scheme. The architecture of the network is shown in Fig. 4.

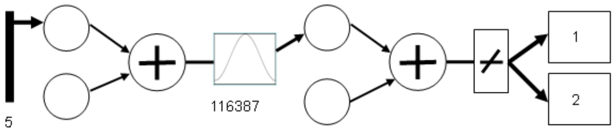

Fig. 4. Architecture of the two output network.

1) Simulation results/performance of proposed RBF ANN

The network is trained using the radial basis function (exact fit) ANN under MATLAB Neural Network tool box. The spread constant is taken as 1.0 in the present work. The trained network is tested with data sets consisting of trained data of 15 data sets (24 hours averaging). The tested results are shown in Table II. These test data sets show the expected output. It is clear from the Table II that the ANN has successfully identified the missing parameter. The output of the ANN for a particular two missing parameter are exactly the same as expected.

TABLE II: RESULTS OF IMPLEMENTED METHOD FOR MISSING TWO PARAMETERS.

Input Parameters (24 Hours Averaging)					Simulated Results of RBFNN	
CO_2 (ppm)	SO_2 (ppm)	NO_2 (ppm)	O_2 (%)	RH (%)	CO (ppm)	T (^oC)
421.6541	0.130	0.025	20.45000	69.40	8.16	31.97
421.7654	0.130	0.020	19.29375	71.00	8.45	32.34
427.2375	0.110	0.150	19.02250	77.00	8.46	30.17
423.8738	0.105	0.026	19.32000	79.90	8.39	29.46
424.3525	0.110	0.025	20.16125	81.90	8.25	28.87
423.3925	0.116	0.020	19.22875	81.10	8.45	29.35
423.3938	0.115	0.025	19.39500	82.40	8.14	28.96
423.3925	0.130	0.029	19.25875	85.20	8.10	28.44
428.6788	0.120	0.025	19.05500	84.40	8.10	29.09
423.8738	0.105	0.026	18.64500	81.10	8.40	29.77
428.6788	0.110	0.025	19.05500	81.40	8.10	30.05
428.6788	0.110	0.025	19.05500	84.30	8.10	28.83
428.6788	0.120	0.025	20.01000	85.90	8.10	28.21
425.7963	0.100	0.021	19.25000	86.70	8.24	27.68
427.2375	0.110	0.140	19.00200	89.10	8.46	27.40

IV. RESULTS AND DISCUSSIONS

In the 1st part of this paper an effort has been made to describe the best estimation techniques for the missing data of completely random nature. Two techniques have been proposed as the solution to estimate these missing data. The two techniques are based on radial basis function neural network and linear regression models. These two techniques are successfully demonstrated in previous sections to predict the indoor environmental parameters. Both these techniques are discussed individually and also compared with each other.

Table I and Table II show the results of implemented scheme for missing one parameter and missing two parameters.

The linear regression analysis technique has been used to evaluate the missing one or two environmental parameters. Application of this technique has resulted in development of seven mathematical relations. To generate the formulae, first 15 days data of the month is used and the developed formulae is validated on next 15 days data of the month. For further analysis and validation of the developed mathematical formulae, those formulae are used which has correlation coefficients greater then 0.98. Higher value of correlation coefficients i.e. greater a 0.98 for the validation represent the high accuracy in predicting the approximate value of missing data.

We found, the RBFNN simulated error varies from 0.249E (-3) % - 0.768E (-3) % for missing one parameter and 0.969E (-3) % – 0.35E (-2) % for missing two parameters.

Further analysis of the models proves that the result obtained by RBFNN_S model is more accurate and suitable for estimating the missing parameters for short duration.

V. Conclusions

This paper has been discuses RBF based ANN technique for the solution of the problem related to missing data for short duration.

Radial basis function (exact fit) approach has been used for ANN training and testing. Further the radial basis neural networks can be designed directly by fitting special response inputs where they carry out the most good. The applicability of the GUI of Neural Network tool box under MATLAB environment has been explored. The technique as described here is successfully used in estimating the missing parameters in both the cases with fairly good accuracy.

Findings also showed that the RBFNN based ANN seem to perform better in cases where the missing data is completely random.

References

[1] V. Tremp, R. Neuneier, and S. Ahmad, "Efficients methods of dealing with missing data in supervised learning," *Advances in Neural Information Processing Systems*, MIT Press, Cambridge MA, vol. 7, 1995.

[2] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art," *Psychological Methods*, vol. 7, no. 2, pp. 147-177, 2002.

[3] F. V. Nelwamondo, S. Mohamed, and T. Marwala, "Missing data: A comparision of neural network and expectation maximization techniques," *Current Science*, vol. 93, no. 11, pp. 1514 - 1521, 10 December 2007.

[4] P. K. Sharpe and R. J. Solly, "Dealing with missing values in neural network based diagnostic systems," *Neural Computing and Applications*, vol. 3, pp. 73-77, 1995.

[5] B. M. Boshkoska and M. Stankovski, "Prediction of missing data for Ozone concentration using support vector machine and radial basis neural networks," *Informatica*, vol. 31, pp. 425-430, 2007.

[6] M. Kolehmainen, H. Martikainen, and J. Ruuskanen, "Neural networks and periodic components used in air quality forecasting," *Atmospheric Environment*, vol. 35, pp. 815-825, 2001.

[7] M. Kolehmainen, H. Martikainen, T. Hiltunen, and J. Ruuskanen, "Forecasting air quality parameters using hybrid neural network modeling," *Environmental Monitoring and Assessment*, vol. 65, pp. 277-286, 2000.

[8] G. Calori, M. Clemente, R. D. Maria, S. Finardi, F. Lollobrigida, and G. Tinarelli, "Air quality integrated modeling in turn urban area," *Environment Modelling and Software*, vol. 21, pp. 468-476, 2006.

[9] C. M. Ennett, M. Frize, and C. R. Walker, "Influence of missing values on artificial neural network performance," *Proceeding of Medinfo 2001*, London, UK, pp. 449-453, Sep. 2-5, 2001.

[10] A. Pelliccioni and T. Tirabassi, "Air dispersion model and neural network: a new perspective for integrated models in the simulation of complex situations," *Environment Modelling and Software*, vol. 21, pp. 539-546, 2006.

[11] E. A. Basurko, G. I. Berastegi, and I. Madariaga, "Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area," *Environment Modelling and Software*, vol. 21, pp. 430-446, 2006.

[12] M. Benghenem and A. Mellit, "Radial basis function network-based prediction of global solar radiation data: Application for sizing of a stand-alone photovoltaic system at Al-Madinah," *Saudi Arabia, Energy*, vol. 35, pp. 3751-3762, 2010.

[13] L. Gao, M. X. Liu, G. X. Sheng, Y. Y. Sui, and Y. K. Zhuang, "Fuzzy discrimination analysis method based on RBFNN and its application in soft measurement," *Proceedings of the IEEE International Conference on Automation and Logistics*, Qinadao, china, pp. 2603-2607, Sep. 2008.

[14] J. Hertz, R. G. Palmer, and A. S. Krogh, "Introduction to the theory of neural computation," *Perseus Books*, 1990, ISBN 0-201-51560-1.

[15] J. Lawrence, "Introduction to neural networks," *California Scientific Software Press.*, 1994, ISBN 1-883157-00-5.

[16] M. Timothy, "Signal and image processing with neural networks," *John Wiley and Sons, Inc.*, 1994, ISBN 0-471-04963-8.

[17] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd Edition, New York: John Wiley, 2000.

[18] J. L. Schafer, *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall, 1997.

[19] S. F. Burk, "Amethod of estimation of missing values in multivariable data suitable for use with an electronic computer," *J. royal Statistic Soc*, pp. 302-306, 1990.

Anuj Kumar received his Ph.D. Degree in Embedded System from Indian Institute of Technology Delhi, India in 2011, M. Tech. degree in Instrumentation from National Institute of Technology Kurukshetra, Haryana, India in 2004 and M. Phil in Instrumentation from Indian Institute of Technology Roorkee, UA, India in 2000. He was with the APL Intelligent Embedded, New Delhi, India where he was involved in development of Microcontroller Based Application from 2000-2002. In 2004, he joined DellSoft Technologies, New Delhi, India as an Instrumentation Engineer – EDA. From 2011 to till date, as a Post doctoral research fellow at Electrical and computers Engineering Department, University of Seoul, South Korea, India. His research Interest includes Smart Sensing System, Intelligent System, Microcontroller Based Applications and Instrumentation Electronics.

Prof. Hiesik Kim was born in 1953 in Kyoung-Ju, Korea. He got the bachelor degree in Mechanical Engineering at Seoul National University in 1977. And master degree in Production Engineering at KAIST (Korea Advanced Institute of Science and Technology) and doctor degree (Doktor-Ingenieur) in Production Engineering (FhG-IPA) at Stuttgart University, Germany, in 1987 by adviser Prof. Dr.-Ing. H.-J. Warnecke. He worked as a technical official at the Ministry of Science and Technology of Korean Government (1979-1982) and then as Senior Researcher, CAD/CAM Research Lab at KIST (Korea Advanced Institute of Science Technology) (1987-1989) and he is now a Professor at the Dep't of Electrical and Computer Engineering of University of Seoul since 1989. He has served as vice dean at the Engineering College. His research areas are optical measurement of geometries, applications of sensors for automation and Image Processing.