On-Line Modelling And Forecasting Of Carbon Monoxide Concentrations Using Hybrid Multilayered Perceptron Network

29872

Usharani Raji[1], M.Y.Mashor[2], A.N.Ali[3], A.F. Sadullah[4]

[1,2,3]*Control and Electronics Intelligent System (CELIS) Research Group, School of E & Ec Engineering*
University Sains Malaysia, Engineering Campus,
14300 Nibong Tebal, Seberang Perai Selatan, Pulau Pinang, Malaysia
Tel:+604-5937788 ext.6093, Email:vsharani@yahoo.com[1], yusof@eng.usm.my[2], nazri@eng.usm.my[3]

[4]*Engineering Innovation & Technology Development Unit,*
University Sains Malaysia, Engineering Campus,
14300 Nibong Tebal, Seberang Perai Selatan, Pulau Pinang, Malaysia
Tel:+604-5937788 ext.5050, Email:cefrhn@eng.usm.my

Abstract

This paper discusses on-line modelling and forecasting of carbon monoxide (CO) concentrations using Hybrid Multilayered Perceptron (HMLP) Network. The HMLP network is trained using Modified Recursive Prediction Error (MRPE) algorithm. In the literature, CO concentrations forecasting are always conducted using off-line modelling. In this preliminary study, on-line modelling is introduced, which means network parameters will be updated for each data sample. Historical CO concentration values are used to provide on-line forecasting. The proposed on-line modelling will be compared to off-line modelling. The performances of those modelling are evaluated using index of coefficient (R^2) and mean square error (MSE). On-line model are found to perform better compared to off-line model.

Keywords:

on-line, real time, off-line, carbon monoxide, forecasting, hybrid multilayered perceptron

Introduction

Carbon Monoxide (CO) is produced from incomplete burning of carbon containing fuels. CO is a colourless, odourless, tasteless and a very poisonous gas. According to the Journal of American Medical Association (JAMA), 1500 people die annually due to accidental carbon monoxide poisoning and 10000 people seek medical attention [1]. Motor vehicles are a major source of CO in urban areas.

A lot of researches have been carried out using different methodology on CO concentrations forecasting. One of the methods was by using univariate linear stochastic models based on Box-Jenkins modelling technique [2]. This model sufficiently needs long historical data set for model formulation. Another approach was by using Box Jenkins transfer function noise model (TFN), the forecast performance were better compared to the first approach [3]. Besides that, Gaussian and regression models were also implemented for CO forecasting [4][5]. Another method for forecasting is implementation of Neural Network (NN). NN have been proved mathematically of representing non-linear systems. A NN known as "Brainmaker" using back propagation algorithm were used to predict CO concentration with an accuracy of R^2=0.69 [6]. Forecasting on other gases using NN were reviewed since not much of studies have been done specifically on CO. In another study, prediction of $PM_{2.5}$ concentrations was carried out using multilayer neural network, linear regression and persistence. The predictions produced by those methods were compared and NN was found to give the best result [7]. In another work, an improved neural network model that combines principal component study and radial basis function (PCA/RBF) were used to forecast hourly time series of RSP, NO_x and NO_2. The improved model was found to perform better than simple RBF [8].

In the present study, CO concentrations forecasting will be performed using Hybrid Multilayered Perceptron Network (HMLP) with Modified Recursive Prediction Error Algorithm (MRPE). The on-line and off-line modelling will be compared using performance evaluation test mentioned above.

Methodology

Neural Network Forecasting

In this study, HMLP network with MRPE algorithm is used to perform CO concentrations forecasting because the network was found to perform better compared to Multilayered Perceptron (MLP), Radial Basis Function (RBF) and Hybrid Radial Basis Function in [9]. The theoretical discussion of HMLP network and MRPE algorithm can be found in [9].

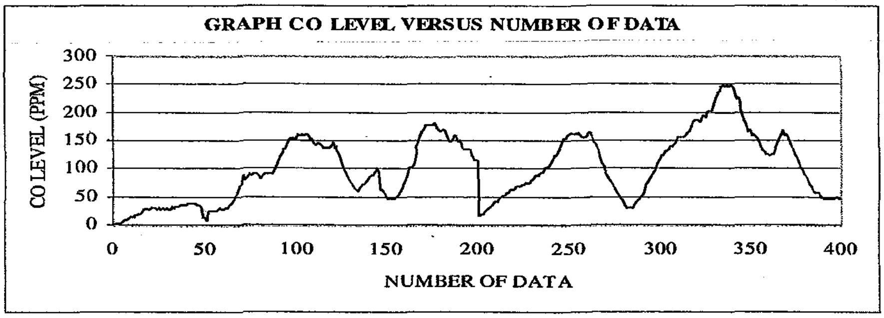

Figure 1 - CO Concentrations versus Number of Data

In this study, HMLP network will perform CO concentrations forecasting using on-line and off-line techniques. Usually neural network forecasting will be carried out using off-line model. In this model, the data set will be divided into two, which consists of training set and testing set. HMLP network parameters will only be updated using the chosen algorithm in the training phase. HMLP network will be trained repeatedly for the training data samples until minimum prediction error is achieved. Then, the final network parameters will be tested using the independent testing set.

In on-line modelling and forecasting, the HMLP network parameters will be updated for each sample of CO measurement. In this case, the network parameters will be always updated to achieve minimum prediction error. For this model, data are not divided such as off-line technique.

In this paper, multiple steps ahead prediction test (MSA) are carried out to obtain the future value of CO concentrations level. The number of steps ahead to be forecasted has been limited to 8. Forecasting performance of high prediction level relies on lower prediction level. If the prediction error for one step ahead is minimum, it will provide a good initial condition for the higher level prediction [10].

Results and Discussion

Data from simulated environment

The simulated environment data set was collected from a simulated environment in a lab. It contains 400 data were collected with sampling time of 10 seconds. The data obtained from simulated environment is shown in Figure 1. The analysis to determine the best input lags are carried out for off-line modelling. It is very important to determine the best input lags in order to produce good results in term of forecasting performance. The analysis is limited to 30 input lags, because further increases of input lags are not found to provide any significant improvement. Number of hidden nodes used are 3, for both on-line and off-line model. For off-line modelling, the first 100 data are used to train the network and the remaining 300 data are used to test the fitted model and to calculate index of coefficient (R^2). The best three inputs lags obtained from off-line modelling are shown in Table 1.

Table 1- Selected Input Lags for Simulated Environment Data

Number	Input Lags
1	(t-1)(t-2)
2	(t-1)(t-2)(t-3)(t-4)
3	(t-1)(t-2)(t-3)(t-4)(t-5)

Off-line model is able to produce good results by using only 1 epoch, the performance degrades by using multiple epochs. R^2 calculated from each set of input lags using 1 epoch is shown in Table 2. It can be seen that R^2 values varies for different set of input lags. It can be noted that model seems to perform well by using Lag 2. It means that in order to obtain good performance, 4 past values of CO concentrations need to be considered to perform CO concentrations forecasting.

Table 2 - R^2 Values Achieved for Off-line Model

Number of Steps	Lag 1	Lag 2	Lag 3
1	0.9728	0.9792	0.9764
2	0.9263	0.9527	0.9438
3	0.8598	0.9172	0.8952
4	0.7734	0.8692	0.8252
5	0.6683	0.8091	0.7297
6	0.5472	0.7391	0.6113
7	0.4122	0.6627	0.4758
8	0.2639	0.5800	0.3216

For on-line modeling, the last 300 data are used to calculate R^2. The analysis to determine the best input lags for on-line modeling are also carried out. For the comparison to be fair, the best three inputs lags obtained from off-line model are used to calculate R^2 for this model.

One input lag is chosen from the analysis performed for on-line model, since it is found to give the best performance compared to the other lags. The four input lags used to calculate R^2 value are shown in Table 3.

The R^2 values achieved for the input lags used is shown in Table 4. First, the results are compared using the same input lags as the off-line model. In that case, Lag 2 gave the best results compared to the others. From the analysis shown in Table 4, it can be concluded that Lag 4 is found to perform even better compared to Lag 2. This means that only 11 past CO concentrations value need to be obtained in order to produce good accuracy in terms of forecasting performance.

Table 3 - Selected Input Lags for Simulated Environment Data

Number	Input Lags
1	(t-1)(t-2)
2	(t-1)(t-2)(t-3)(t-4)
3	(t-1)(t-2)(t-3)(t-4)(t-5)
4	(t-1)(t-2)(t-3)(t-4)(t-5)....(t-11)

Table 4 - R^2 Values Achieved for On-line Model

Number of Steps	Lag 1	Lag 2	Lag 3	Lag 4
1	0.9764	0.9800	0.9791	0.9810
2	0.9387	0.9550	0.9509	0.9552
3	0.8854	0.9190	0.9146	0.9195
4	0.8171	0.8700	0.8661	0.8701
5	0.7352	0.8100	0.8071	0.8100
6	0.6430	0.7360	0.7400	0.7447
7	0.5437	0.6658	0.6671	0.6756
8	0.4384	0.5931	0.5892	0.5997

The MSE plot for the whole set of data is shown in Figure 2, which indicates that HMLP network parameters converge rapidly. The MSE plot converge around 220 data samples.

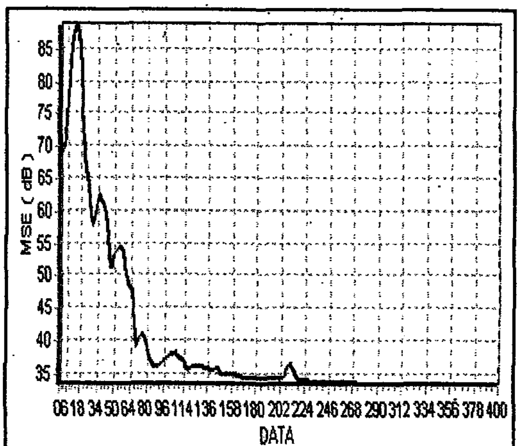

Figure 2 - MSE for Simulated Environment Data

From the results shown using simulated environment data, it can be concluded that both on-line and off-line model using HMLP network produce good forecasting performance in terms of R^2 test. For comparison using the best input lags obtained from off-line model, it can be concluded that on-line model is found to perform slightly better compared to off-line model. Besides that, it performs even better by using the best input lags obtained from the analysis using on-line model. The differences between both models become more noticeable when the number of steps ahead for forecasting is increased. The maximum differences achieved by R^2 value using the same lags are around 0.2 and 0.3.

In this section, both models are trained using 100 data samples only, in order to show performance differences between both models when the networks are under trained. From the results obtained, it is proven that on-line model is able to perform well even when the network is trained using small amount of data. Off-line model does not give good prediction for multiple steps ahead forecasting. It can be concluded that the number of training data used has a very strong influence towards the off-line model performance.

Industrial Data

The industrial data set was obtained from Malaysian Environmental Department (ASMA). These data contains the average hourly of CO measurements. These data were obtained from industrial area in Penang. 1000 data samples are used to perform the task for this paper. The Industrial Data is shown in Figure 3. 1000 data are used to perform CO forecasting, it contains average hourly CO concentrations measurement. For off-line modeling, the first 600 data are used to train the network while the remaining 400 data are used to test the data samples. The testing data samples are used to calculate R^2. The analysis to determine the best input lags are performed in order to obtain the best results from the model. The analysis is limited to 30 input lags because more input lags do not give any significant improvement in terms of R^2 test. The best three inputs lags obtained from off-line modelling are shown in Table 5.

Table 5 - Selected Input Lags for Industrial Data

Number	Input Lags
1	(t-1)(t-2)
2	(t-1)(t-2)(t-3)(t-4)(t-5)
3	(t-1)(t-2)(t-3)(t-4)(t-5)(t-6)

R^2 values calculated for each set of inputs are shown in Table 6. For industrial data, the off-line models are trained using 3 epochs, because the model does not produce good results by using 1 epoch. The R^2 values achieved with 1 epoch using the selected input lags are shown in Table 6, while the results obtained using 3 epochs are shown in Table 7.

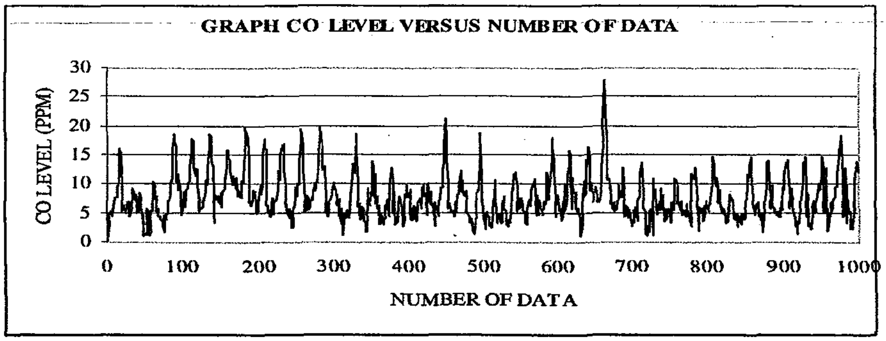

Figure 5 - CO Concentrations versus Number of Data

Table 6. R^2 Values Achieved for Off-line Model Using 1 Epoch

Number of Steps	Lag 1	Lag 2	Lag 3
1	0.7072	0.6840	0.6587
2	0.3972	0.3944	0.2968
3	0.1541	0.2067	-0.0222
4	-0.0218	0.0620	-0.2781
5	-0.1683	-0.0297	-0.5005
6	-0.2581	-0.0739	-0.6526
7	-0.3271	-0.0971	-0.8317
8	-0.3570	-0.1044	-0.9703

Table 7- R^2 Values Achieved for Off-line Model Using 3 Epochs

Number of Steps	Lag 1	Lag 2	Lag 3
1	0.7251	0.7168	0.7174
2	0.4464	0.4392	0.4182
3	0.2423	0.2707	0.1762
4	0.0958	0.1717	-0.0077
5	0.0189	0.1237	-0.172
6	-0.0940	0.1052	-0.2777
7	-0.1427	0.0954	-0.3741
8	-0.1745	0.0872	-0.4618

From the results shown in Table 6 and Table 7, it can be noted that off-line model using 3 epochs is found to perform better compared to single epoch. Off-line model using Lag 2 gave better results compared to the others. This means that 5 past CO concentrations value need to be considered in order to produce better accuracy of forecasting performance.

For on-line modelling, the last 400 data are used to calculate R^2 value. Analysis to determine the best input lags for on-line modeling are carried out. For the comparison to be fair, the best three inputs lags obtained from off-line model are used to calculate R^2 value. One best input lag is chosen from analysis performed for on-line model, since it is found to perform better compared to the input lags obtained from off-line model. The four input lags used to calculate R^2 value are shown in Table 8. The R^2 value achieved for the input lags used is shown in Table 9.

Table 8 - Input Lags used for Industrial Data

Number	Input Lags
1	(t-1)(t-2)
2	(t-1)(t-2)(t-3)(t-4)(t-5)
3	(t-1)(t-2)(t-3)(t-4)(t-5)(t-6)
4	(t-1)(t-2)(t-3)(t-4)(t-5)....(t-21)

First, the results are compared using the same input lags with off-line model. In that case, Lag 2 gave the best results compared to the others. From the analysis shown in Table 9, it can be concluded that Lag 4 is found to perform even better than Lag 2. This means 21 past CO concentrations value need to be considered in order to produce better accuracy in terms of forecasting performance compared to others.

Table 9. R^2 Values Achieved for On-line Model

Number of Steps	Lag 1	Lag 2	Lag 3	Lag 4
1	0.7354	0.7300	0.7165	0.7337
2	0.4638	0.4558	0.4239	0.5152
3	0.2685	0.2637	0.1122	0.3855
4	0.1367	0.1379	0.0544	0.3071
5	0.0286	0.2637	0.1122	0.2387
6	-0.054	0.0128	-0.176	0.1951
7	-0.123	-0.063	-0.250	0.183
8	-0.171	-0.025	-0.313	0.1809

The MSE plot for 1000 data samples is shown in Figure 4. The MSE converges to an acceptable value after about 500 data samples. It indicates that HMLP network requires 500 data samples for off-line modelling.

It can be concluded that the R^2 values achieved by on-line and off-line models are only good for one step ahead forecasting. The R^2 values obtained are found to drop drastically for two steps ahead forecasting with the differences between 0.2 and 0.3. Overall, both, models do not give good prediction after 3 steps. This could be due to under sampling of CO concentration measurements, since the data obtained from ASMA were sampled hourly. Besides that, Figure 5 shows that CO concentrations level are very dynamic because the plot shows that CO level fluctuates heavily. The on-line model gives better results compared to off-line model for one and two step ahead forecasting. The differences between both the models are more noticeable when the number of forecasting step increased. On-line model performs better even though the off-line model used for comparison applied multiple epochs. For industrial data, on-line model is found to perform better compared to off-line model.

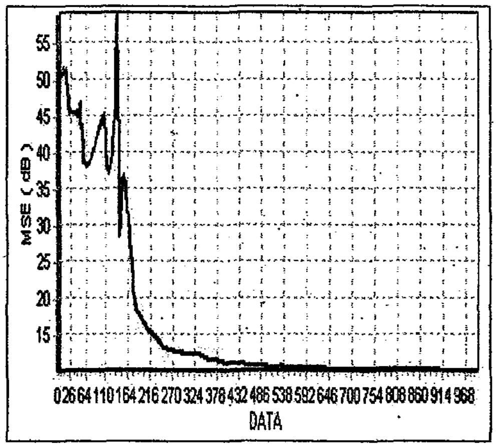

Figure 4 - MSE Plot for Industrial Data

Conclusions

This study proves that HMLP network using Modified Recursive Prediction Error Algorithm (MRPE) can be used to perform CO concentrations forecasting. Both on-line and off-line models provide good results in terms of forecasting performance. However, on-line model is found to perform better compared to off-line model using both sets of data due to its learning capability. Besides that, sampling time used has significant contribution towards forecasting performance. If the sampling time is very large, the network will not be able to produce good results for multi steps ahead forecasting. Overall, forecasting performance depends on the dynamic of the data set.

References

[1] Carbon Monoxide Alert, 2001. Citing Internet Sources URL http:// www.homesafe.com/coalert, 2001.

[2] Sharma, P., Khare, M., 2000. Real- time Prediction of Extreme Ambient carbon Monoxide Concentrations due to Vehicular Exhaust Emissions using Univariate Linear Stochastic Models. Transportation Research: Part D, Vol.5, No.1:59-69.

[3] Sharma, P., Khare, M., 2001. Short-term, Real- time Prediction of Extreme Ambient carbon Monoxide Concentrations due to Vehicular Exhaust Emissions using Transfer Function Noise Model, Transportation Research: Part D, Vol.6, No.2:141-146.

[4] Zickus, M. & Kvietkus, K., 1998. Urban Air Pollution Forecast based on the Gaussian and Regression Models. In Proceedings of Sixth International Conference on Air Pollution, 515-523.

[5] Comrie, A.C. & Diem, J.E., 1999. Climatology and Forecast Modelling in Phoenix, Arizona. Atmospheric Environment, Vol.33, No.30:5023-5036.

[6] Moseholm, L., Silva, J. & Larson, T., 1996. Forecasting Carbon Monoxide Carbon Monoxide Concentrations Near a Sheltered Intersection using Video Traffic surveillance and Neural Networks. Transportation Research D, Vol. 1, No. 1: 15-28.

[7] Patricio, P., Trier, A. & Reyes, J., 2000. Prediction of $PM_{2.5}$ Concentrations Several Hours in Advance using Neural networks in Santiago, Chile. Atmospheric Environment, Vol.34, No.8:1189-1196.

[8] Lu, W.Z., Wang, W.J., Wang, X.K., Yan, S.H. & Lam, J.C., 2004. Potential Assessment of a Neural Network Model with PCA/RBF approach for Forecasting Pollutant Trends in Mong Kok urban air, Hong Kong. Environmental Research, Vol.96, No.1:79-87.

[9] Mashor, M.Y., 2000. Hybrid Multilayered Perceptron Networks, International Journal of System Science, Vol.31:771-785.

[10]Berger, C.S., 1995. Recursive Single Layer Nets for Output Error Dynamic Models, IEEE Transactions On Neural Networks, Vol.5: 508-511.