Processes Modeling using Artificial Neural Network in the Presence of Oultiers

고영철 [1], 박화규 [1], 봉복준 [1], 손주찬 [1], 왕지남 [2]
[1] 시스템공학연구소 시스템통합연구부
[2] 아주대학교 산업공학과

Abstract

Ouliers, unexpected extraordinary observations that look discordant from most observation in a data set are commonplace in various kinds of data analysis. Since the effect of ouliers on model identification could be serious, the aim of this paper is to present some ways of handling outliers in given data set and to specify a model in the presence of outliers. A procedure based on neural network which identifies outliers, removes their effects, and specifies a model for the underlying process is proposed. In contrast with traditional parametric methods requiring to estimate the model's structure and parameters before detecting outliers, the proposed procedure is a non-parametric method without the estimation of model's structure and parameters before handling outliers and could be applied for real problems in the presence of oultiers. The proposed methodology is performed as followings. Firstly, outliers are detected and the detected outliers replace the prediction values using outliers detection neural network. The data set removing the effect of outliers is retraining using neural network. Therefore the effects of outliers are removed and the modeling precision can be improved. Experimental results show that the proposed method is suitable for predicting data set in the presence of outliers.

Introduction

Computer-controlled manufacturing systems require process monitoring and control which could be performed effectively by an on-line tracking task. Tracking manufacturing processes are required to identify the model structure and parameters of the current underlying process. When the underlying processes is changing dynamically, it would be a difficult task to keep track of the process model structure and parameters. Also, it is difficult to analyze the data representing system's state in the case of existing unexpected extraordinary observations in a data set. Generally, the unexpected extraordinary observations which are apart from most data are called outliers.

When you gather data in real world, outliers are commonplace. Roughly speaking, there are three sources of outliers[4]. First, the distribution of the model's random disturbances often has longer tails than the normal distribution, resulting in a greatly increased chance of larger disturbance. Second, the data set may contain erroneous values. Erroneous values can result form misinterpreted questions, incorrectly recorded answers, keypunch errors, etc. Third the model itself, typically linear in (transformations of) the variables, is only an approximation to reality. It is apt to be a poor representation of the process generating the data of extreme values of the explanatory variables.

This paper handles outliers in a given data set. If outliers are contained in a data set , the results of analysis are distortion. So, before data analysis such as modeling, estimation and so on, it is important to be able to identify these outliers and remove their effects from the sample. This study is to propose the method handling outliers. The proposed method is essentially based on M. Johnson's method using forecast error to detect outliers the neural network to remove their effects on data analysis.

The paper is organized as follows. Section 2 is a brief statement of the problem. That section describes the problem to be covered in this paper. Section3 discusses the basic principle of the proposed model identification using neural network which detects the outliers and removes their effects in data analysis. Section 4 presents experimental results of computer simulation that verify the proposed approach. In the

final section, some discussion and concluding remarks are given.

Statement of the Problem

Observations which are separated from the rest of the data are called outliers. These aberrant observations have an effect on the result of modeling. If you may detect outliers in the sample, it needs to be appropriately treatment for them. In the modeling of systems, it is difficult to identify a model structure and parameters where the underlying process is non-stationary. The situation also arises when outliers are contained in the target data of analysis. In this case, the effects of outliers on model identification are very serious and outliers have an influence on modeling accuracy of the underlying systems. Therefore, they must be detected before modeling and the detected outliers must be removed in the data set.

In traditional parametric methods, analysts must find model's structure and parameters before detecting and removing outliers. If they can't find its structure and parameters, it is impossible to handle outliers. Traditional parameter methods are as fallowed. Tasy[7] handled the time series model in presence of outliers based on the iterative estimation of Chang and Tiao and the extended sample autocorrelation function(ESACF) model identification method of Tasy and Tiao. Kim, Bae and Lee presented handling outliers in a given data and investigated the effect of the analysis result before and after outliers reject[5]. Chang, Tiao and Chen proposed the estimation of time series parameters using autoregressive-integrated-moving-average in the presence of outliers [3] and so on.

The main objective of this research is to develop a handling outliers for model identification in the presence of outliers. The proposed method using neural network is non-parametric method that it needn't find model's structure and parameters in order to detect and remove outliers in contrast with traditional parametric methods. First, the outliers-detection module is detecting outliers using difference between original values and prediction values using neural network. After detecting outliers, the outliers-elimination module replaces the detectied outliers with the prediction values using neural network and their effects on data analysis could be eliminated

The Procedure of Model Identification in Presence of Outliers

The procedure of modeling in presence of outliers is performed as followed. The procedure consists of outliers-detection module, outliers-elimination module. That module detects outliers in a given data set and this module replaces them with thier prediction values using neural network. After eliminating outliers, the retraining of the sample using neural network could remove their effects on data analysis.

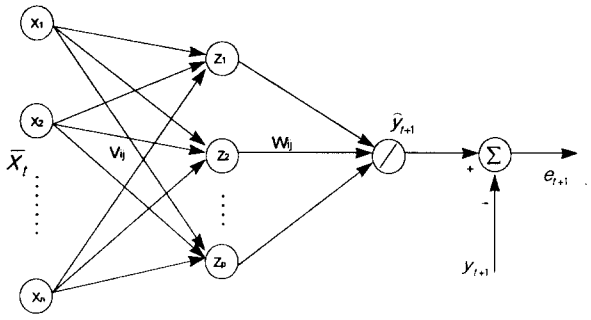
In order to detect outliers in a given data set and replace them with the appropriate values, this paper proposes the outliers-detection module using the forecast error and the outliers-elimination module using the neural network's prediction values. First, the outliers-detection module is performed as followed. In this paper, the outliers-detection is based on the method of M. Johnson which is using the forecast error to detect outliers[6]. In order to calculate forecast error, this paper uses the one-step ahead prediction values, $\bar{x}_T(T-1)$, using neural network. Because of that, it is not necessary to know the model structure and parameters before handling outliers.

Outliers can be identified by analyzing the forecast error, $e_1(T) = x_T - \bar{x}_T(T-1)$. If this error is large, it may be concluded that the observations x_T came from a different process. The test for outliers might logically take the form.

$$\left|\frac{e_1(T)}{\hat{\sigma}_e}\right| > K \qquad (1)$$

where K is 4 or 5. If the inequality holds, x_T is considered an outliers.

After detecting outliers through the outliers-detection module, the detected outliers must be replaced with another appropriate values and its effect on modeling could be eliminated. Fig. 1 shows the architecture of neural network. Each input consists of 10 consecutive observations.



$\bar{X}_t = (X_t, X_{t-1}, X_{t-1}, \ldots\ldots, X_{t-9})$

Fig. 1 : Architecture of Outliers-Detection module using neural network

The detected outliers must be substituted by the prediction values using neural network. Accordingly, modeling accuracy could be improved because outliers, the cause of an adverse effect on modeling, are eliminated.

$$y_T \leftarrow \tilde{y}_T \qquad (2)$$

In the case of analyzing data which contains outliers, it is in difficulties for data analysis in the cause of outliers and the precision of data analysis could be extremely decreased. This study proposes the method which enhances modeling accuracy removing the effect of outliers in the sample.

Experimental Results

A computer simulation is presented for validation of proposed methodology. This implementation is to investigate how well the proposed method can detect the outliers and remove its effects. The outliers-detection module is evaluated by how correctly outliers are detected when it is necessary. The outliers-elimination module is evaluated by modeling accuracy. One-step ahead prediction errors are analyzed for evaluation of modeling accuracy

Two different time series are used for implementation. The first series shown in the Table 1 is based on a synthetically generated linear series. The second series, which is a non-linear stepwise changing series, is generated synthetically based on the paper Chen *et al*[1]. From Table 2, the model parameters and structure of the second series is presented and here u_t denotes uniform distribution from 0 to 1.

Table 1. Sequence of Linear Stationary Input Series with Model Parameters

Model	Number of Inputs	Model Structure
AR(2)	200	$y_t = 1.49y_{t-1} - 0.653y_{t-2} + e_t$
AR(3)	200	$y_t = 2.146y_{t-1} - 1.598y_{t-2} + 0.409y_{t-3} + e_t$
AR(4)	200	$y_t = 1.876y_{t-1} - 1.781y_{t-2} - 1.20y_{t-} - 0.373y_{t-4} + e_t$
AR(5)	200	$y_t = 1.840y_{t-1} - 0.893y_{t-2} - 0.613y_{t-3} - 0.879y_{t-4} - 0.350y_{t-5} + e_t$

Table 2. Sequence of Nonlinear Stationary Input Series with Model Parameters

Model	Number of Inputs	Model Structure
NLX	200	$y_t = 0.757y_{t-1} + 0.389u_{t-1} - 0.037y_{t-1}^2 + 0.379y_{t-1}u_{t-1} + 0.063u_{t-1}^3 - 0.739e_{t-1} - 0.368u_{t-1}e_{t-1} + e_t$

Each input consists of 10 observation. Moving block of fixed width of 10 observations is used for the design of input. For a given model, 200 inputs which are 2000 observations, are used for simulation. Outliers are randomly generated by computer simulation. Its range is from $\pm 3\sigma$ to $\pm 6\sigma$. Mean Square Error(MSE) is emploued as a modeling performance measure. The mathematical expression of MSE can be described as

$$MSE = \frac{\sum_{t=1}^{N} (e_t)^2}{N} \qquad (3)$$

where e_t denotes an one-step ahead prediction error of a given technique at time t and N is the total number of evaluation.

Three different techniques are compared with the simulation results. The 'No Outliers' shown Table

3. denotes the MSE of one-step ahead prediction errors of the clean data which doesn't contain outliers. The 'Eliminating Outliers' in Table 3. means the MSE of one-step ahead prediction errors of the model after eliminating outliers. The 'Non-Elimination' in Table 3. means the MSE of one-step ahead prediction errors of the data without removing outliers.

The simulation result shows that the proposed method has as good as the performance for modeling the linear series and the non-linear series comparing with 'No outliers'. The performance of the proposed method is more excellent than that of 'No Outliers'. This implies that the proposed method could be applied for modeling identification in presence of outliers.

Table 3. Comparison of modeling performance(Linear series)

Model Structure	No Outliers	Eliminating Outliers	Non-Elimination
	MSE	MSE	MSE
AR(2)	0.01521	0.01561	1.26731
AR(3)	0.01592	0.01616	1.56732
AR(4)	0.01553	0.01593	1.45191
AR(5)	0.01532	0.01548	1.78450

Table 3. Comparison of modeling performance(Non-linear series)

Model Structure	No Outliers	Eliminating Outliers	Non-Elimination
	MSE	MSE	MSE
Non-linear	0.01802	0.01819	2.0081

Conclusion

A handling outliers scheme is proposed for model identification in the presence of outliers. The proposed method plays on important role in detecting and eliminating outliers in very serious case in the presence of outliers. Using the proposed scheme, The performance of modeling in the presence of outliers could be improved.

Traditional parametric method should find out the model structure and parameters in order to detect and remove outliers. But the proposed method is non-parametric method which is not necessary to estimate the model structure and parameters before handling outliers and could directly be applied for real problems in the presence of oultiers. For example, machine conditions could be diagnosed by analyzing the obtained data representing the machine's state using this method.

This research considers outliers as meaningless values. But outliers could sometimes provide useful information. Therefore, future research might identify its meaning and appropriately handle them.

Reference

[1] C. F. N. Chen, S. A. Billings and P. M. Grant. A parallel recursive prediction error algorithm for training layered neural networks, *Int. J. Control*, Vol 51, pp1215~1228, 1990

[2] Gi-Nam Wang and Young Cheol Go, On-line Neuro-Tracking of Non-Stationary Manufacturing Processes, *Computers and Industrial Engineering* Vol. 30, No. 3, pp449~461, 1996

[3] I. Chang, G. C. Tiao and C. Chen, Estimation of Time Series Parameters in the presence of outliers. *Technometrics*, Vol. 30, No. 2, pp193~204, May 1988

[4] J. Eatwell, M. Melgate and P. Newman, *The New Palgrave Time Series and Statistics*, Norton & Company, 1990

[5] K. S. Kim, Y. J. Bae and J. G. Lee. The Effect of Outliers in Regression Analysis, Korea Quality Management, Vol 24, No. 3, pp158~171, April 1996

[6] M. Johnson, *Forecasting and Time Series Analysis*, McGrawHill, 1976

[7]R. S. Tasy, Time Series Model Specification in the Presence of Outliers. *Journal of American Statistical Association*, Vol. 81, No. 393, pp132~141, March 1986