



Outliers detection in environmental monitoring databases

Hugo Garces*, Daniel Sbarbaro

Department of Electrical Engineering, University of Concepción, Casilla 43-C, Correo 3, Concepción, Chile

ARTICLE INFO

Article history:

Received 23 November 2009

Received in revised form

6 October 2010

Accepted 19 October 2010

Available online 20 November 2010

Keywords:

Data preprocessing

Neural networks

Nonlinear PLS

Outlier identification

ABSTRACT

Environmental monitoring is nowadays an important task in many industrial operations. In order to comply with strong environmental laws, they have implemented monitoring systems based on a network of air quality and meteorological stations providing real-time measurements of key variables associated to the distribution of pollutants in surrounding areas. These measurements can be contaminated by outliers, which must be discarded in order to have a consistent set of data. This work presents a nonlinear procedure for outliers detection based on residual analysis of regression with Partial Least Squares and Artificial Neural Networks. In order to minimize the negative effect of outliers in the training dataset a learning algorithm with regularization is proposed. This algorithm is based on a Quasi-Newton optimization method and it was tested on a simulated nonlinear process, on real data from environmental monitoring contaminated with synthetic outliers, and finally applied to a real environmental monitoring data obtained from a monitoring station and having natural outliers. The results are encouraging and further developments are foreseen for including information from neighboring stations and emission source operation.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Online environmental monitoring is an important task associated to the operation of industrial plants, since these measurements are used to minimize the impact of pollutants in surrounding areas. This impact can be assessed by measuring a range of substances considered contaminants such as sulphur dioxide (SO₂), nitrogen dioxide (NO₂), ozone (O₃), particulate material (PM₁₀ or PM_{2.5}), etc. All of them are measured in µg/m³ or ppb by chromatographs. These substances in dangerous concentrations can cause multiples health problems in the population, for instance high levels of SO₂ can cause respiratory diseases such as bronchitis, pulmonary edema, and even heart attack. High levels of NO₂ can cause from skin irritation to severe lung damage. To carry out environmental monitoring a network of air quality stations, located inside and outside the industrial complex, is normally deployed. These stations measure the concentrations of contaminant substances and meteorological stations provide measurements of wind speed, wind direction, temperature, and humidity. Usually the air quality stations consider data loggers that acquire the measurements obtained by chromatographs and send them to an environmental database hosted in a master server, which can be used by clients to get and analyze the data. Unfortunately, these measurements can have outliers generated by specific abnormal

operation conditions in the emission source, bad calibrated or faulty instruments, electrical faults, problems in the data logger or communication channels. Outliers do not have a formal definition, but many authors define them as measurements not consistent with most of the dataset. Their presence has a masking effect if the outliers are considered as normal measurements and they are not removed from the dataset. On the other hand, if a measurement is wrongly labelled as an outlier, then it represents a swamping effect. Masking effect implies contaminated dataset with measurements that do not represent normal system relationships and swamping effect involves a dataset that is not representative of the full system relationships. Pearson (2002) analyzes the detection of outliers in datasets used for system identification, and he shows that outliers are not always associated to failures in measurements or data collection systems, they can also be a product of physical effects representing patterns mainly related to abnormal system operation. Outliers can produce many negative effects in the estimated values if they are not taken into account in the analysis of the dataset, as it has been reported for instance in Pearson (2002), Kadlec et al. (2009), and Warne et al. (2004).

Outliers detections have been applied in many fields such as chemical engineering and data mining. In spectroscopy, for instance, Wiegand et al. (2009) present simultaneous variable selection and outliers detection based on robust genetic algorithm. Bao and Dai (2009) propose an iterative procedure for outliers detection based on robust scaling of Partial Least Square (PLS) to predict gasoline properties. Lin et al. (2007) present outlier detection in virtual sensor design for chemical processes based on unidimensional Hampel identifier and multidimensional Principal

* Corresponding author. Tel.: +56 41 2204247.

E-mail addresses: hugarces@udec.cl, hugarces@gmail.com (H. Garces), dsbarbar@udec.cl (D. Sbarbaro).

Component Analysis (PCA) approach. In a more general setting, [Warne et al. \(2004\)](#) describe two fundamental tasks in inferential measurement design: data collection and influential variable selection. In this context, the method proposed by [Jolliffe \(1986\)](#) has been used to detect outliers. [Chiang et al. \(2003\)](#) compare a set of outlier detection algorithms with simulated data from an industrial process and also show the effect of multiple outliers in auto scaling and robust scaling of the dataset. Their results outperform conventional outlier detection methods based on Principal Components Analysis (PCA). [Fortuna et al. \(2007\)](#) also consider outlier detection based on Jolliffe's method within their methodology for designing industrial virtual sensors. All these references show the importance outlier detection in any data-driven task, such as empirical modelling or virtual sensor design based on experimental data.

Partial Least Square (PLS) has been extensively used for outlier detection in dataset with linear features ([Jolliffe, 1986](#); [Fortuna et al., 2007](#)). In order to deal with dataset having nonlinear features, [Baffi et al. \(1999a\)](#) have developed a nonlinear approach for the inner model in PLS. The input weights are updated by an error based procedure considering a quadratic relation between input and output scores called Quadratic Partial Least Squares (QPLS). This approach has been tested using nonlinear datasets demonstrating better performance than conventional linear PLS in estimation tasks. The same authors have also developed a second nonlinear approach using Multilayer Perceptron (MLP) neural networks for the inner model in PLS [Baffi et al. \(1999b\)](#) (MLPPLS). They have demonstrated that an error based update procedure for input weights improves performance estimation compared to linear PLS, QPLS, and MLPPLS without input weights update. [Bang et al. \(2003\)](#) propose a Takagi–Sugeno–Kang (TSK) fuzzy model for the inner model, this approach is closely related to the work carried out by [Baffi et al. \(1999b\)](#) using radial basis functions (RBF). The main advantage of fuzzy inner model with respect to “black box” models is its interpretability. A fuzzy model provides the possibility of using expert's knowledge in order to design the inner structure design process. This Fuzzy PLS regression method, including both input weights update and constant input weights, has been tested using simulated dataset and spectral measurements from diesel fuels. The results demonstrate that this method can outperform other PLS approaches such as LPLS and QPLS. More recently, [Li et al. \(2007\)](#) have presented an approach based on the work of [Baffi et al. \(1999b\)](#) to estimate the residual life of a large generator stator insulation. The nonlinear modelling task was decomposed into a set of linear outer relation and a simple nonlinear inner relation, which were performed by a number of single input–single output models. The results outperformed linear regression tools such as multivariate linear regression and linear PLS.

This work considers a MLP neural network inner model, since environmental data usually present nonlinear features and they are not clustered. The training algorithm for estimating the parameters of MLP plays an important role in the final performance attained by the network. The presence of outliers precludes the use of standard error-backpropagation algorithm. Outliers in training datasets usually produce large differences between the real measurement and the estimated one, which leads to standard methods to provide biased estimated. Different approaches have been proposed to mitigate the effect of outliers on the estimated parameters. [Zhao et al. \(2004\)](#) have proposed the use of a weighted error-back-propagation algorithm in order to deal with training dataset contaminated with outliers. Their method weight the error according to a scalar detection index in order to attenuate the effects of outliers. Since this method does not remove the outliers from the training dataset, it can be used with small dataset. However, the detection index is not robust when a large number of outliers are present. [Liano \(1996\)](#) analyzes the mean square error behaviour,

normally used as minimization index in neural networks training methods, and shows that outliers increase the estimation error deviating the parameters from the real values. In order to lessen the effect of outliers, he has proposed a method based on maximum likelihood estimators and a Cauchy error distribution, which has heavier tails respect to Gaussian distribution. Thus, if there are outliers in the dataset, then the error is kept bounded and hence parameters update is more stable. The main disadvantage is its slow convergence, because during the initial training iterations gross errors are mainly due to model initialization. [Chen and Jain \(1994\)](#) present a modified neural networks training algorithm, which is robust against outliers, by defining a time variant minimization index based on robust estimators. This time variant index uses some knowledge about the underlying nonlinear function obtained from a MLP model to reject samples with gross error and reduce their effect over the parameters update. This approach also attenuates parameter update in the initial stages of training due to the initial lack of knowledge of the underlying function. [Chuang et al. \(2000\)](#), however, propose a solution to the problem of robust neural networks training by using a time variant outlier detection test. The detection threshold of the detection tests is adjusted while the model is trained, since at the beginning the function to be identified is unknown and the error may not provide a good indication to identify outliers. On the other hand, after several training epochs the model has learned the unknown function and the error can be used to discriminate outliers. This method outperforms other robust neural networks training methods with noise-free, small noise and gross error model datasets, but performance presents strong dependence on threshold initialization and the definition of an epoch dependent function.

In this work, instead of using the standard Levenberg–Marquand method proposed by [Baffi et al. \(1999b\)](#) and the robust training method previously described, a simple and improved method based on Quasi-Newton method with a regularization term is proposed. The regularization term improves generalization of input/output relation, gives a smooth parameters update and provides a robust estimation, if outliers are present in the training set ([Bishop, 1997](#)), without the drawbacks of the previously described methods.

This paper is organized as follows: Section 2 describes methods for outlier detection and Section 3 describes the proposed method for training the inner nonlinear structure of a PLS model. In Section 4 some synthetic examples are used to demonstrate the performance of the proposed algorithm. In addition, some experiment using real data are also presented. Finally, in Section 5 some conclusions and future work are outlined.

2. Methods for outliers detection

Outlier detection methods can be classified into unsupervised and supervised ones. Unsupervised distance based techniques calculate a scalar value representing how far a particular measurement is from the centre (or reference) of the data considering parameters such as location and dispersion. Supervised projection techniques, usually based on Partial Least Squares or Principal Components Regression, project the original dataset through a model into a set of latent variables where outliers became apparent by applying a function over projected dataset. Distance based approaches are sensitive to location and dispersion parameters and the value of the threshold distance to detect outliers based on a particular distribution. Projection methods do not assume a data distribution, but they are sensitive to model structure representing the nonlinear patterns contained in the original dataset and projected into the latent variables. These methods can also be applied to analyze high dimensional dataset such as geochemical or chemometrics, where the dimensions range in the thousands

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات