Chemometrics and Intelligent Laboratory Systems 111 (2012) 53–58

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Removal of the effects of outliers in batch process data through maximum correntropy estimator

Jose Co Munoz [a,b], Junghui Chen [a,*]

[a] R&D Center for Membrane Technology, Department of Chemical Engineering, Chung-Yuan Christian University, Chung-Li, Taiwan, 320, Republic of China
[b] Department of Chemical Engineering, University of Philippines, 1101 Diliman, Quezon City, Philippines

ARTICLE INFO

Article history:
Received 12 May 2011
Received in revised form 18 October 2011
Accepted 17 November 2011
Available online 27 November 2011

Keywords:
Correntropy
Data rectification
Dispersion
Outlier data
Wavelet analysis

ABSTRACT

In some batch processes, the control of a variable of interest is done by controlling another variable. The relationship between the variable of interest and the controlled variable is established by an empirical process model which should be built from reliable data. The presence of outliers in the variable of interest affects the reliability of the data which can result in erroneous interpretations concerning the variable of interest. In this paper, a novel method, referred to as the maximum correntropy estimator (MCE), of removing the effects of outliers by the use of a robust estimator that maximizes correntropy is proposed. The effectiveness of MCE is dependent on the proper selection of kernel width which is a parameter that specifies the minimum magnitude of error arising from an outlier which will be excluded by the estimator in approximating the trend of the majority of the values of the variable of interest with time. An initial set of outliers is determined by a preliminary detection method and from these outliers, the minimum value of error is estimated for the kernel width. The trend which is an unknown function is fitted by a linear combination of scaling functions of a given resolution and its coefficients are determined by MCE. Given the unknown function, the remaining outliers are identified on a proposed error cut-off value. The values of all the outliers are replaced with estimates from the unknown function so that the batch data set is now free of outliers and can then be used in building a reliable process model. The advantages of the proposed method, data from a chemical batch reactor, are presented to help readers delve into the matter.

1. Introduction

In batch processes, the use of outlier-free data is very important for building reliable input–output models for controllers and running processes efficiently. Outliers are incorrect measurements unrelated to the actual output values, so they affect the analysis of the data sets containing them. Basically, they are observations that appear to be inconsistent with the remainder of the data set [1]. They are found infrequently in most data sets but their absence cannot always be presumed. Outliers come from various sources. In an automatic data acquisition environment, they arise from sporadic malfunctioning of sensors and equipment. In batch polymerization, they are generated from incorrectly recorded viscosity data usually obtained from a tedious laboratory procedure. In anaerobic digestion, clusters of outliers appear because of recorded results of unsuccessful batch runs.

Minimizing the effects of outliers in a data set always begins with an assessment of the data set to determine the presence of outliers. There are many outlier detection methods cited in literature and all of these methods have shown to be able to detect obvious outliers, which are the isolated measurements deviating largely from the trend exhibited by the majority of the data values. In many cases, however, the methods could not detect all the inconspicuous outliers as they are masked by their adjacent outliers. It is very likely that the data set obtained after performing an outlier detection method still contains outliers. The trend exhibited by the majority of the values of the data set over time within a batch can be used as a tool for outlier identification. The trend is approximated by a linear combination of basis functions [2–6] whose coefficients are determined by using robust estimators. The accuracy of the approximation is affected by the values of the estimator parameters [7,8]. Any data value at a particular time deviating significantly from the trend is considered an outlier. If the trend is accurately determined, all the outliers are correctly identified and their values can be replaced with calculated values from the trend.

In this paper, a robust estimator that maximizes correntropy is proposed to handle outliers in a data set. This robust estimator is referred to as the maximum correntropy estimator (MCE). MCE uses correntropy defined as the mean of a Gaussian function of the difference between two random variables [9]. Since the Gaussian function approaches zero for large error values, this measure of goodness of

* Corresponding author. Fax: +886 3 2654199.
E-mail address: jason@wavenet.cycu.edu.tw (J. Chen).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.11.007

fit is robust to the effects of large errors arising from outliers. The effectiveness of MCE is controlled by the value of a parameter called the kernel width. From the literature review, it is unclear how the value of the kernel width is specified. In this paper, a formula is also proposed for the kernel width computation. The proposed formula requires an estimate of the minimum magnitude of an error arising from an outlier which is identified by the proposed preliminary outlier detection method.

The remainder of this paper is organized as follows. In Section 2, some outlier identification and detection techniques, especially the proposed preliminary outlier detection method, are discussed. In Section 3, the use of MCE in determining the coefficients of basis functions and the proposed computation of the kernel width are presented. In Section 4, the effectiveness of MCE in removing outliers from a batch reactor data set is demonstrated. Lastly, the conclusion is given in Section 5.

2. Preliminary outlier detection methods

Most outlier detection methods in chemical engineering extract dynamic data associated with normal operating conditions from data obtained from continuous processes, not from batch processes. The methods are applied in the area of dynamic data reconciliation which is a process of adjusting or reconciling the time varying process measurements to obtain more accurate estimates of real data such as flow rates and temperatures that are consistent with material and energy balances. In differentiating outliers from normal data values, the distance criterion is used to determine a weighting factor which is assigned to a data value and which determines the influence of the error arising from the same data value on the data reconciliation procedure. Chen et al. [10] proposed that for each measured data value, the distance is expressed in terms of minimum distance which is defined as the minimum absolute difference of a specific value of the variable of interest with every other value in the data set. A data value is considered as an outlier if it has a minimum distance larger than twice the average minimum distance. The weighting factor assigned to a normal data value is one and that of an outlier is less than one as it is the ratio of twice the average minimum distance to the minimum distance of the outlier. In Chen's method, when outliers have values that are close to each other, their minimum distances are smaller than twice the average minimum distance and their weighting factor is one. As such, the outliers are incorrectly regarded as normal data values. To eliminate this effect where an outlier masks the presence of another outlier, Abu-el-zeet et al. [11] propose that the distance of a data value be expressed in terms of the absolute value of the difference between the data value and the mean of all the data values of the variable of interest, not in terms of the minimum distances of data values to each other. Applied to a data set where the range of variation of the data values is large, the distance of the normal data values at the start or end of the batch may exceed twice the average distance of all the data values so these normal data values are mistakenly identified as outliers and this is referred to as swamping effect.

A more complicated scheme of outlier detection is proposed by Castano and Kunoth [12]. Their method compares two different functions estimated from two data sets: one with a suspect outlier point and another one without. A suspect outlier is considered a true outlier if there is a significant difference in local energy of the two functions. Their method can detect outliers constituting a maximum 5% of the total data but it is no longer effective for a data set containing a higher percentage of outliers.

In this paper, Chen's method [10] is modified to include time t in the list of output variables to be analyzed for matching the dynamic batch operation. Time t is referred to the time when the measurements are taken and the distance calculation is modified to include the contribution of the additional variable t. The minimum distance between a data point $(t_i,\ y_i)$ and all the other data points $(t_j,\ y_j)$ $j \neq i$ is calculated using the following equation

$$\underset{\min,i}{\mathrm{DIST}} = \min_{\text{all } j \neq i} \left[\left(t_i - t_j\right)^2 + \left(y_i - y_j\right)^2 \right]^{\frac{1}{2}}. \tag{1}$$

With the inclusion of t in Eq. (1), an outlier has significantly larger minimum distance than the normal data points since the normal data points usually appear in clusters. In cases when the outliers occur in cluster, the method can still identify an individual outlier since their minimum distances will be very different from one another. The proposed formulation will make the minimum distances of the outliers much larger than those of the normal data points, so the outliers will lie on the right end of a minimum distance distribution. This minimum distance distribution can be represented by a histogram whose data points with extremely large minimum distances are identified as outliers. Since low quality data may consist of as high as 10% outliers, the distribution is assumed to have arbitrarily a conservative value of 5% outliers. In view of these considerations, this paper proposes that an outlier is a data point whose minimum distance of an outlier is greater than the minimum distance corresponding to the 95 percentile. The identified outliers are used to estimate the minimum magnitude of the errors arising from an outlier and they are removed from the data set before the trend exhibited by the majority of data points is determined.

3. MCE outlier identification

MCE uses correntropy which is a generalized similarity measure between two random variables, Y and W. Mathematically, correntropy is expressed as

$$\mathrm{V}_{\sigma_{kw}}(Y,W) = E\left[k_{\sigma_{kw}}(Y-W)\right] \tag{2}$$

where $E[\cdot]$ is the mathematical expectation operator, $k_{\sigma_{kw}}(Y-W)$ is a Gaussian kernel function of realizations $(y_i - w_i)$ of the random variable $Y-W$. The Gaussian kernel function $k_{\sigma_{kw}}(y_i - w_i)$ is defined as

$$k_{\sigma_{kw}}(y_i - w_i) = \frac{1}{\sigma_{kw}\sqrt{2\pi}} \exp\left(-\frac{(y_i - w_i)^2}{2\sigma_{kw}^2} \right) \tag{3}$$

and σ_{kw} is the kernel width. Usually the joint probability distribution of the random variables Y and W is unknown and only a finite number of data $\{(y_i, w_i)\}_{i=1}^{N}$ is available, so the sample estimator of correntropy is defined to be

$$\hat{V}_{N,\sigma_{kw}}(Y,W) = \frac{1}{N}\sum_{N}^{1} k_{\sigma_{kw}}(y_i - w_i). \tag{4}$$

Since the effect of the two random variables comes in the form of their difference, correntropy can be expressed in terms of an error random variable $E = Y - W$, and the sample estimator of correntropy can be expressed as

$$\hat{V}_{N,\sigma_{kw}}(E) = \frac{1}{N}\sum_{i=1}^{N} k_{\sigma_{kw}}(e_i) = \frac{1}{\sigma_{kw}\sqrt{2\pi}} \frac{1}{N} \sum_{i=1}^{N} \exp\left(-\left(\frac{e_i}{\sigma_{kw}\sqrt{2}} \right)^2 \right). \tag{5}$$

Thus, in a batch process, given data samples $\{(t_i, y_i)\}_{i=1}^{N}$ and mathematical expression of the unknown function $f(t)$ which approximates the dependence of y_i on t_i, correntropy can be used as a measure to describe how well $f(t)$ fits the data provided an appropriate value of the kernel width is used. As the errors between the data y_i and the predicted values $f(t_i)$ decrease to a minimum, correntropy monotonically increases to a maximum indicating that the best functional approximation has been achieved. When the model is