Chinese Journal of Chemical Engineering 22 (2014) 820–827

Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

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

Process Monitor

# Adaptive Local Outlier Probability for Dynamic Process Monitoring ☆

Yuxin Ma, Hongbo Shi *, Mengling Wang

*Key Laboratory of Advanced Control and Optimization for Chemical Processes of Ministry of Education, East China University of Science and Technology, Shanghai 200237, China*

ARTICLE INFO

*Article history:*
Received 24 December 2013
Received in revised form 28 January 2014
Accepted 7 February 2014
Available online 19 June 2014

*Keywords:*
Time-varying
Complex data distribution
Local outlier probability
Multi-mode
Fault detection

ABSTRACT

Complex industrial processes often have multiple operating modes and present time-varying behavior. The data in one mode may follow specific Gaussian or non-Gaussian distributions. In this paper, a numerically efficient moving window local outlier probability algorithm is proposed. Its key feature is the capability to handle complex data distributions and incursive operating condition changes including slow dynamic variations and instant mode shifts. First, a two-step adaption approach is introduced and some designed updating rules are applied to keep the monitoring model up-to-date. Then, a semi-supervised monitoring strategy is developed with an updating switch rule to deal with mode changes. Based on local probability models, the algorithm has a superior ability in detecting faulty conditions and fast adapting to slow variations and new operating modes. Finally, the utility of the proposed method is demonstrated with a numerical example and a non-isothermal continuous stirred tank reactor.



## 1. Introduction

In industrial processes, operating conditions are usually affected by some slow variations denoted as time-varying characteristics, caused by some dynamic behavior such as seasonal fluctuation, catalyst deactivation, equipment aging, sensor or process drifting, preventive maintenance and cleaning [1]. Generally, effects of the time-varying behavior on the mean and covariance of variables cannot be neglected, so there may be many false alarms if conventional multivariate statistical process monitoring (MSPM) methods are applied directly [2]. In order to maintain process efficiency for a long period of time, numerous adaptive methods have been developed. Recursive MSPM methods and methods based on the moving window strategy are two alternative widely used approaches [3,4].

Multimodality is another important feature of industrial processes due to changes of market demands, alternations of feedstock or variations of manufacturing strategy. The difference between the characteristics of nearby operating conditions is always significant, so intensive studies have been carried out with either multiple local models or a single global model [5,6]. While it is more practical to accommodate the time-varying behavior and multimode features together. The developed methods can be divided into two categories. One is the adaptive clustering methods. Teppola *et al.* [7] applied adaptive fuzzy C-means algorithms on the score values of principle component analysis (PCA) to monitor a wastewater treatment plant. Liu [8] used an adaptive Takagi-Sugeno fuzzy model on PCA subspace to model a large scale nonlinear system containing many operating regions. Since PCA is used as a preprocessing tool, monitoring results of these two methods more or less depend on and be restricted by the capability of PCA. Petković *et al.* [9] designed an on-line adaptive clustering method utilizing a generalized information potential. Although previously unseen functioning modes can be included by introducing an adaptive expert system, the method suffers from a non negligible detection delay. The other category is adaptive statistical methods. Improved recursive algorithms based on recursive PCA or the signed digraph were proposed with some if-then rules to distinguish process condition changes from disturbances [10–12]. Ge and Song [13] introduced the just-in-time-learning strategy to the modeling procedure of local least squares support vector regression and the residuals between the real output and the predicted one was analyzed by a two-step information extraction strategy. Xie and Shi [14] and Yu [15] developed two different dynamic fashions of Gaussian mixture model (GMM) separately based on the moving window strategy and a particle filter re-sampling method.

The problem of complex data distributions in time-varying and multimode processes has scarcely been addressed. Although the moving window strategy has been proven to be effective, it still encounters some limitations when incorporated with statistical methods such as

☆ Supported by the National Natural Science Foundation of China (61374140), Shanghai Postdoctoral Sustentation Fund (12R21412600), the Fundamental Research Funds for the Central Universities (WH1214039), and Shanghai Pujiang Program (12PJ1402200).

* Corresponding author.
*E-mail address:* hbshi@ecust.edu.cn (H. Shi).

http://dx.doi.org/10.1016/j.cjche.2014.05.015

PCA, partial least squares (PLS) or GMM. Since the variables of an industrial process may satisfy specified Gaussian or non-Gaussian distributions, and high order statistics are usually helpful to reveal more information from the data [16–18], adaptive monitoring algorithms should be developed, which can explore both Gaussianity and non-Gaussianity of process data. Local outlier probability (LoOP) is an unsupervised data mining technique proposed for outlier detection [19]. It combines the idea of local, density-based outlier scoring with a probabilistic, statistically-oriented approach, and assigns the probability of being an outlier to all data records. Since a normalization procedure is included, LoOP is independent of any specific data distribution. Therefore, a combination of LoOP and moving window strategy should be potential to tackle these problems.

The main contribution of this paper is to propose a numerically efficient moving window LoOP algorithm for monitoring industrial processes with complex data distributions, time-varying property and multiple operating modes. Some designed rules are introduced and incorporated with a two-step adaption approach to ensure that the monitoring model can be updated at a high speed. To cope with the multimode features, a semi-supervised monitoring strategy is employed, and an update termination rule is developed to prevent the monitoring model contaminated by faults or disturbances. Since the method is based on local probabilistic models, the accuracy of model is higher and it will be much easier to detect faulty conditions.

## 2. Adaptive Process Monitoring Based on Moving Window Loop

For low computation burden and practical applications, it is fast and reasonable to only update the information of those samples whose neighbors have changed due to the insertion and discard of samples. The key problems addressed in this section are how to find the affected samples and how to update their information.

### 2.1. Offline initialization

To make an initialization and calculate the LoOP value for each sample $\boldsymbol{x}_j$ ($j = 1, 2, \ldots, L$) with dimension $D$ in the initial window $\boldsymbol{W}_1$, its $k$ nearest neighbors are found as follows, with its neighborhood set in $\boldsymbol{W}_1$ can be recognized as $\mathbf{knn}_1(\boldsymbol{x}_j)$.

$$d\left(\boldsymbol{x}_j,\boldsymbol{x}_p\right)=\sqrt{\sum_{n=1}^{D}\left|x_{jn}-x_{pn}\right|^2}\qquad\left(p\neq j\ \text{ and }\ \boldsymbol{x}_j,\boldsymbol{x}_p\in\boldsymbol{W}_1\right)\tag{1}$$

Assuming that samples in $\mathbf{knn}_1(\boldsymbol{x}_j)$ are centered around $\boldsymbol{x}_j$, then we can define probabilistic set distance as:

$$pdist_1\left(\boldsymbol{x}_j\right)=\lambda\cdot\sqrt{\textstyle\sum_{\boldsymbol{x}_p\in\mathbf{knn}_1(\boldsymbol{x}_j)}d\left(\boldsymbol{x}_j,\boldsymbol{x}_p\right)^2/k}\tag{2}$$

where λ is a weighted factor usually taken as 2. For estimating the density around $\boldsymbol{x}_j$, the probabilistic local outlier factor (PLOF) is defined as follows with function $E(.)$ used to compute the expectation of PLOF in the current window.

$$PLOF_1\left(\boldsymbol{x}_j\right)=pdist_1\left(\boldsymbol{x}_j\right)/\left(E_{\boldsymbol{x}_p\in\mathbf{knn}_1(\boldsymbol{x}_j)}\left[pdist_1\left(\boldsymbol{x}_p\right)\right]\right)-1\tag{3}$$

To achieve normalization, the aggregate value $nPLOF_1$ which can be considered as a standard deviation of PLOF values is obtained:

$$nPLOF_1=\lambda\cdot\sqrt{E\left[(PLOF_1)^2\right]}\tag{4}$$

Finally, by applying the Gaussian error function, the local outlier probability indicating the probability that a sample is an outlier can be calculated as:

$$LoOP_1\left(\boldsymbol{x}_j\right)=\max\left\{0,\operatorname{erf}\left(PLOF_1\left(\boldsymbol{x}_j\right)\Big/\left(\sqrt{2}\cdot nPLOF_1\right)\right)\right\}\tag{5}$$

where erf(.) is the Gaussian error function applied to obtain a probabilistic value.

### 2.2. Online updating and process monitoring

By applying the moving window strategy, a two-step adaption procedure is introduced to update the monitoring model. Some more details of the adaption procedure for a window size $L$ are as follows.

Step 1: discard

The effect of eliminating the oldest sample $\boldsymbol{x}_i$ from the previous window $\boldsymbol{W}_i$ on the mean and variance can be evaluated as follows.

$$\tilde{\boldsymbol{\mu}}=(L\boldsymbol{\mu}_i-\boldsymbol{x}_i)/(L-1)\tag{6}$$

$$\Delta\tilde{\boldsymbol{\mu}}=\boldsymbol{\mu}_i-\tilde{\boldsymbol{\mu}}\tag{7}$$

$$\tilde{\boldsymbol{\sigma}}(m)^2=\frac{1}{L-2}\left((L-1)\cdot\left([\boldsymbol{\sigma}_i(m)]^2-\left[\Delta\tilde{\boldsymbol{\mu}}(m)\right]^2\right)-[\boldsymbol{x}_i(m)-\boldsymbol{\mu}_i(m)]^2\right)\quad(m=1,2\ldots,D)\tag{8}$$

$$\tilde{\boldsymbol{\Sigma}}=\operatorname{diag}\left[\tilde{\boldsymbol{\sigma}}(1),\tilde{\boldsymbol{\sigma}}(2),\cdots,\tilde{\boldsymbol{\sigma}}(D)\right]\tag{9}$$

where diag(.) is the function used to calculate the diagonal matrix. Eq. (6) describes the updating of the variable mean while Eqs. (7)–(9) describe the updating of the variable variance.

After moving all the information about $\boldsymbol{x}_i$ from the current monitoring model, a set $\boldsymbol{S}_{i-1}^0(i>1)$ is constructed to store the samples, in which $\boldsymbol{x}_i$ is one of their $k$ nearest neighbors.

$$\boldsymbol{S}_{i-1}^0=\boldsymbol{S}_{i-1}^0\cup\left\{\boldsymbol{x}_j\right\},\ \text{ if }\ i\neq j,\ \ \boldsymbol{x}_j\in\boldsymbol{W}_i\ \text{ and }\ \boldsymbol{x}_i\in\mathbf{knn}_i\left(\boldsymbol{x}_j\right)\tag{10}$$

where $\mathbf{knn}_i(\boldsymbol{x}_j)$ represents the neighborhood set of sample $\boldsymbol{x}_j$ in window $\boldsymbol{W}_i$. Obviously, if $\boldsymbol{x}_j\in\boldsymbol{S}_{i-1}^0$, due to the deletion of $\boldsymbol{x}_i$, the neighborhood set $\mathbf{knn}_i(\boldsymbol{x}_j)$ will change.

Step 2: insertion

When a new sample $\boldsymbol{x}_{i+L}$ is judged normal and added into the data matrix, the updated mean vector and variance in $\boldsymbol{W}_{i+1}$ are computed as follows.

$$\boldsymbol{\mu}_{i+1}=\left[(L-1)\tilde{\boldsymbol{\mu}}+\boldsymbol{x}_{i+L}\right]/L\tag{11}$$

$$\Delta\boldsymbol{\mu}_{i+1}=\boldsymbol{\mu}_{i+1}-\tilde{\boldsymbol{\mu}}\tag{12}$$

$$\boldsymbol{\sigma}_{i+1}(m)^2=\frac{1}{L-1}\left((L-2)\cdot\left[\tilde{\boldsymbol{\sigma}}(m)\right]^2+(L-1)\cdot\left[\Delta\boldsymbol{\mu}_{i+1}(m)\right]^2+\left[\boldsymbol{x}_{i+L}(m)-\boldsymbol{\mu}_{i+1}(m)\right]^2\right)\quad(m=1,2\ldots,D)\tag{13}$$

$$\boldsymbol{\Sigma}_{i+1}=\operatorname{diag}\left[\boldsymbol{\sigma}_{i+1}(1),\boldsymbol{\sigma}_{i+1}(2),\cdots,\boldsymbol{\sigma}_{i+1}(D)\right]\tag{14}$$

Eq. (11) describes the updating of the mean vector while Eqs. (12)–(14) describe the updating of the variance. However, only for those with new sample $\boldsymbol{x}_{i+L}$ among their $k$ nearest