2021 IEEE 3rd International Conference on Communications, Information System and Computer Engineering (CISCE 2021)

A multimodal analysis method based on border-peeling clustering for soft sensor

Ziyan Jiang
Department of Industrial Engineering and Operations Research
University of California, Berkeley
CA 94720, USA
Email: duoerxs@163.com

Qiuyue Zhang
Department of Information Technology
Beijing University of Technology
Beijing, 100124, China

Bin Chen
Jiluan College
Nanchang University
Jiangxi, 330000, China

***Abstract*—Soft sensors have been widely used in industrial process monitoring. The core of monitoring complicated industrial processes is to recognize multi-modes and strategically apply different sub-model. This paper proposes a new soft sensor-modeling method based on BPC (Border-peeling Clustering) and PLSR (Partial Least Square Regression). Moreover, BPC is robust towards noise and sample unbalanced problems. By iteratively peeling off layers of points, the cores of the latent clusters are revealed, which indicates the different operation modes. The three-phase flow process case proves the effectiveness and superiority of the proposed method. Experimental results of three-phase flow show that the mean square error of the proposed method is 34.1%, which is better than other methods.**

Keywords-Border-Peeling clustering; multimodal; soft sensor; industrial process;

I. INTRODUCTION

In the modern industrial process, it is necessary to measure some key variables in real-time, significant for optimization and monitoring. However, not all variables can be directly measured. Due to the complexity of industrial process products, it is difficult to find a suitable sensor for analysis. The development of a particular sensor often takes a considerable amount of time, incurs additional overhead.

Generally, soft sensors can be divided into two types, mechanism-driven, and data-driven soft sensor methods. Mechanism-driven soft sensors require tremendous expert knowledge. However, a comprehensive understanding of the mechanism is difficult for complex industrial processes. As a result, with the rapid development of artificial intelligence and Internet of Things technology, data-driven soft sensor methods attract unprecedented attention due to the high accuracy and low cost.

The basic theory of data-driven soft sensor methods, such as Principal Component Regression (PCR), PLSR, is to estimate the value of key variables that are difficult to measure by searching for the relationship among these key variables and other easily-measured variables in industrial processes [2-4]. Nevertheless, PCR and PLSR are weak at nonlinear regression. In this situation, some researchers pay attention to Support Vector Regression (SVMR) when industrial processes are producing nonlinear data [5]. Generally, these methods reduce the cost of gaining expert knowledge and purchasing expensive hardware counterparts but give a clear description of an industrial process.

However, the methods mentioned above focus on a single industrial processing mode. Nevertheless, with state-of-the-art technology in industrial processes and automation, industrial manufacture is designed to be multifunctional and multimode [6-9]. In recent, several models have been devoted to multimode industrial processes. Gaussian Mixed Model (GMM) linearly combines several Gaussian distribution functions to fit industrial data from different modes. It is troubling for a soft sensor system to check if data fit Gaussian distribution and indicate the number of modes in a complex industrial process.

To construct efficient and precise soft sensors according to the characters of the modern industrial process, some multimodal modeling strategies based on clustering algorithms are proposed. Generally, a soft sensor system based on clustering contains these steps: (1) Distinguish different clusters from off-line data with a clustering algorithm. (2) Select the suitable method to calculate target variables for every cluster. (3) When received a new data point on-line, estimate a certain cluster that the new data belongs to. (4) Output a predictive value of target variables. This system gains a mode number parameter with the clustering algorithm's help and needs not repeatedly construct prediction models once the off-line model is established. Then, what is concerned is the clustering algorithm. In general, K-means, DBSCAN, and Mean-shift are frequently-used clustering methods. However, these methods do not consider the problem of sample imbalance.

In this paper, a multimodal soft sensor strategy based on Border-Peeling Clustering (BPC) [10] is proposed to solve the unbalanced number of samples of different modes. The peeling process can adapt to the local densities and characteristics to successfully separate adjacent clusters, which improves the accuracy of mode partition.

978-0-7381-1215-2/21/$31.00 ©2021 IEEE

2021 International Conference on Communications, Information System and Computer Engineering (CISCE) | 978-1-6654-0352-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/CISCE52179.2021.9446040

II. REVIEW OF BORDER-PEELING CLUSTERING

Given a dataset $\mathrm{X}=\left(x_1,x_1,\ldots,x_n\right)$ contains n samples and m measured variables. Moreover, $\mathrm{X}^{(t)}$ is a set of points that are unpeeled by the start of t^{th} the iteration. The steps of BPC are listed as follows.:

Positioning Figures and Tables: The Euclidean distance between two samples is $dist(x_i,x_j)$ where $i,j=1,2,\ldots,n$. For each point $x_i\in\mathrm{X}^{(t)}$, denote the set of k nearest neighbors $\mathrm{N}_k^{(t)}\left(x_i\right)$. Furthermore, the reverse k nearest neighbors x_i are given $\mathrm{RN}_k^{(t)}\left(x_i\right)=\{x_j\mid x_i\in\mathrm{N}_k^{(t)}\left(x_j\right)\}$. To estimate distances between points, a pairwise relationship function is introduced as:

$$f\left(x_i,x_j\right)=exp\left(-\frac{dist\left(x_i,x_j\right)^2}{\sigma_j^2}\right) \tag{1}$$

Border peeling: With function f, a density influence $b_i^{(t)}$ of point x_i in t^{th} an iteration is calculated as followed:

$$b_i^{(t)}=\sum_{x_j\in\mathrm{RN}_k^{(t)}(x_i)}f\left(x_i,x_j\right) \tag{2}$$

A classification function $B_i^{(t)}$ is introduced as:

$$B_i^{(t)}=\begin{cases}1,b_i^{(t)}\le\tau^{(t)}\\0,otherwise\end{cases} \tag{3}$$

At the end of t^{th} an iteration, the border points are given by $\mathrm{X}_B^{(t)}=\left\{x_i\in\mathrm{X}^{(t)}\ \&\ B_i^{(t)}=1\right\}$, and $\mathrm{X}^{(t+1)}=\mathrm{X}^{(t)}\setminus\mathrm{X}_B^{(t)}$.

Association: Association between border points and non-border points: After finishing the border point identification in t^{th} iteration, each border point $x_i\in\mathrm{X}_B^{(t)}$ will be associated to a most closed non-border point $\rho_i\in\mathrm{X}^{(t+1)}$. ρ_i is given by

$$\rho_i=\begin{cases}x_j,dist\left(x_i,x_j\right)\le l_i\\\varnothing,dist\left(x_i,x_j\right)>l_i\end{cases} \tag{4}$$

Where $\varnothing$ means x_i is an outlier. l_i is determined at the time the point is classified as a border point.

Iteration: Repeat steps 2) and 3) with $\mathrm{X}^{(t+1)}$ until $t=T$.

Core points merging: After T iteration, the non-border points $\mathrm{X}^{(T+1)}$ are core points, which are clustered by merging close reachable neighborhoods of points. Given core point x_i and x_j, they are reachable when:

$$dist\left(x_i,x_j\right)\le max\left(l_i^{(T)},l_j^{(T)}\right) \tag{5}$$

The merging process is iterative until none of the core points in the two sets are reachable.

Clustering: The points which do not belong to outliers or core points are classed by association and linkage to core points.

III. SOFT SENSOR BASED ON BPC AND PLS

A. Off-line model construction

The first step is to cluster off-line data. Suppose there is a dataset $\mathrm{X}=\left(x_1,x_2,\ldots,x_n\right)$ containing n samples and m measured variables. Dataset X is the initial $\mathrm{X}^{(0)}$ in BPC. Then the BPC starts to border-peeling process.

For a $\mathrm{X}^{(t)}(0\le t\le T)$, the $\mathrm{N}_k^{(t)}\left(x_i\right)$ and $\mathrm{RN}_k^{(t)}\left(x_i\right)$ are calculated for every data point in $\mathrm{X}^{(t)}$, as is defined in Part 2. The $f\left(x_i,x_j\right)$ is calculated as:

$$f\left(x_i,x_j\right)=exp\left(-\frac{dist\left(x_i,x_j\right)^2}{\sigma_j^2}\right) \tag{6}$$

where $\sigma_j=dist\left(x_j,\mathrm{N}_k^{(t)}\left(x_j\right)[k]\right)$ and $dist\left(x_i,x_j\right)=\sqrt[2]{\left(x_{i1}-x_{j1}\right)^2+\ldots+\left(x_{im}-x_{jm}\right)^2}$, which presents the influence level that x_i has on x_j. σ_j takes into account the distribution around x_j. So the density influence factor of every point can be calculated as:

$$b_i^{(t)}=\sum_{x_j\in\mathrm{RN}_k^{(t)}(x_i)}f\left(x_i,x_j\right) \tag{7}$$

According to the factor, points are divided into border points and non-border points by:

$$B_i^{(t)}=\begin{cases}1,b_i^{(t)}\le\tau^{(t)}\\0,otherwise\end{cases} \tag{8}$$

The border points are given by:

$$\mathrm{X}_B^{(t)}=\left\{x_i\in\mathrm{X}^{(t)}\ \&\ B_i^{(t)}=1\right\} \tag{9}$$

and non-border points is defined as

$$\mathrm{X}^{(t+1)}=\left\{x_i\in\mathrm{X}^{(t)}\ \&\ B_i^{(t)}=0\right\} \tag{10}$$

Which is the input of the next iteration $t+1$. In this iteration, border points will be attached to a non-border point. A radius l_i is calculated for each point as:

$$l_i^{(t)}=min\left(\frac{C}{k}\sum_{x_j\in\mathrm{NN}_{B,k}^{(t)}(x_i)}dist\left(x_i,x_j\right),\lambda\right) \tag{11}$$

where $\mathrm{NN}_{B,k}^{(t)}\left(x_i\right)=\bigcup_{r=1}^{t}\mathrm{X}_B^{(r)}$. λ parameter is a parameter serving as the maximal threshold value, and C determines the strictness of the threshold value. If there are no points in $l_i^{(t)}$ the area around x_i, it is labeled as outliers (including all points

attached x_i). Otherwise x_i is attached to the closed non-border point in this area, which is defined as:

$$x_i \to x_j \tag{12}$$

where $x_j \in \mathrm{X}^{(t+1)}$. After finishing the attachment, the next iteration starts.

After T iterations, the remaining non-border points are core points. Suppose there are initially r core points, and every single point represents a set. If there are two points x_i and x_j from two sets and $dist\left(x_i, x_j\right) \le max\left(l_i^{(T)}, l_j^{(T)}\right)$, then the two sets are reachable for each other and will be merged. This iterative process will end until there are no such two reachable sets. In the end, there are P sets of core points $\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_P$. Then, all points except outliers are classified by their linkage to a certain core point. A point x_i will be classified to $\mathrm{C}_p\left(p = 1, 2, \ldots, P\right)$ if there are attachments that satisfy:

$$x_i \to \ldots \to x_j \tag{13}$$

where $x_j \in \mathrm{C}_p$.

The outliers are discarded, and all points in the dataset are classified into P sets named $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_P$. The variable that needs to be predicted is $\mathrm{Y} = \left(y_1, y_2, \ldots, y_n\right)$ which has a one-to-one correspondence with X . Then a PLS model is constructed based on different data clusters. Suppose there is a cluster $\mathrm{X}_p = \left\{x_{p_1}, x_{p_2}, \ldots, x_{p_q}\right\}$ that contains q points, where $\mathrm{Y}_p = \left\{y_{p_1}, y_{p_2}, \ldots, y_{p_q}\right\}$ .The matrixes constructed by X_p and Y_p are $\mathrm{E}_0 = \begin{matrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{q1} & \cdots & x_{qm} \end{matrix}$ and $\mathrm{F}_0 = \begin{matrix} y_1 \\ \vdots \\ y_q \end{matrix}$.

1) $t_1 = \mathrm{E}_0 \mathrm{W}_1$ is a linear combination of E_0 . By maximizing the covariance $\mathrm{cov}\left(t_1, \mathrm{F}_0\right)$, an optimal solution $t_1 = t_1$ is acquired.

2) Linear regression $\begin{cases} \mathrm{E}_0 = \bar{t}_1 \alpha_1^T + \mathrm{E}_1 \\ \mathrm{F}_0 = \bar{t}_1 \beta_1^T + \mathrm{F}_1 \end{cases}$ is performed on E_0 and F_0 , in which the α_1^T and β_1^T are calculated by OLS, and E_1, F_1 are residual.

*3)*Replace E_0 and F_0 with E_1 and F_1, and go to step 1) until the model is precise enough.

*4)*Finally, the PLS model is constructed as:

$$\begin{cases} \mathrm{E}_0 = \bar{t}_1 \alpha_1^T + \ldots + \bar{t}_r \alpha_r^T + \mathrm{E}_r \\ \mathrm{F}_0 = \bar{t}_1 \beta_1^T + \ldots + \bar{t}_r \beta_r^T + \mathrm{F}_r \end{cases} \tag{14}$$

*5)*Every $t_k = w_{k1}^* x_1 + \ldots + w_{km}^* x_m, \left(k = 1, 2, \ldots, r\right)$. Replace all t_k in

$$\mathrm{Y} = t_1 \beta_1 + \ldots + t_r \beta_r \tag{15}$$

then the final PLS model is constructed as follow:

$$\mathrm{Y} = a_1 x_1 + \ldots + a_m x_m \tag{16}$$

*6)*Repeat step (1) to (4) in every single cluster, and there are P PLS models established as follow:

$$\mathrm{Y} = a_{p1} x_1 + \ldots + a_{pm} x_m \left(p = 1, 2, \ldots, P\right) \tag{17}$$

B. Online monitoring

After BPC allied on the dataset, points in the dataset can be divided into outliers, border points, and core points. Core points represent the center of a certain class, and their linkage classifies border points to core points, while outliers are points that too far away from other borders or core points.

Based on steps in BPC, core points have relatively high-density influence due to the border-peeling mechanism. Additionally, core points always stay in the center of a cluster because peeling always happens in the outer sphere. Therefore, characters of a cluster can be represented well by core points, which helps classify new data points from the online industrial process.

According to the off-line model construction part, after T iteration in BPC, there are P sets of core points $\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_P$. When the online system receives a new data point x_{new} , an average distance between the new and core points will be calculated. Assume there are q core points in C_p . The average distance is defined as follow:

$$AveDistance_p = \frac{1}{q}\left(\sum_{x \in \mathrm{C}_p} dist\left(x_{new}, x\right)\right) \tag{18}$$

Then the cluster that the new point belongs to is calculated by:

$$\bar{p} = argmin\left(AveDistance_p\right) \tag{19}$$

According to the off-line modeling part, there are P PLS models constructed with every cluster correspondingly. After detecting the cluster that a new point belongs to, the relative PLS model is applied to predict the target variable. Then the value waiting to be predicted is calculated by:

$$\overline{\mathrm{Y}_{new}} = a_{\bar{p}1} x_1 + \ldots + a_{\bar{p}m} x_m \tag{20}$$

where $x_{new} = \left(x_1, \ldots, x_m\right)$

IV. STUDY CASE

In this section, the effectiveness and superiority of the proposed algorithm are verified by a three-phase flow process.

Here we use 10 variables to predict the flow rate (Variable 11).
The 11 variables used are listed in Table 1.

Table 1. THE USED VARIABLES IN THREE-PHASE FLOW PROCESS

Variable nr	Location	Measured magnitude	Unit
1	PT312	Air delivery pressure	MPa
2	PT401	Pressure in the bottom of the riser	MPa
3	PT408	Pressure on top of the riser	MPa
4	PT403	Pressure in the top separator	MPa
5	PT501	Pressure in 3 phase separator	MPa
6	PT408	Diff. pressure (PT401-PT408)	MPa
7	PT403	Differential pressure over VC404	MPa
8	FT305	Flow rate input air	Sm^3/s
9	FT104	Flow rate input water	kg/s
10	FT407	Flow rate top riser	kg/s
11	FT406	Flow rate top separator output	kg/s

The three-phase flow consists of water, air, and oil mixed and then separated by a device. The working conditions of three-phase flow are mainly determined by water flow and air velocity. When the flow of water and air is faster, the whole process is in a more intense operation. Here, we selected the first 2000 data for analysis and modeling and selected 20% of the data as the verification set. The effect of BPC is shown in Fig. 1, and the working condition of the three-phase flow is shown in Fig.2. It can be seen that in the first 2000 samples, the data are basically in three different working conditions, and there is the problem of unbalanced samples. The data are also grouped into three categories, which illustrates the rationality of clustering.

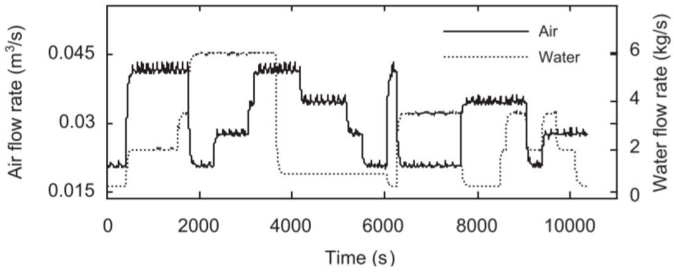

Fig.1 Operational conditions for training data sets T1

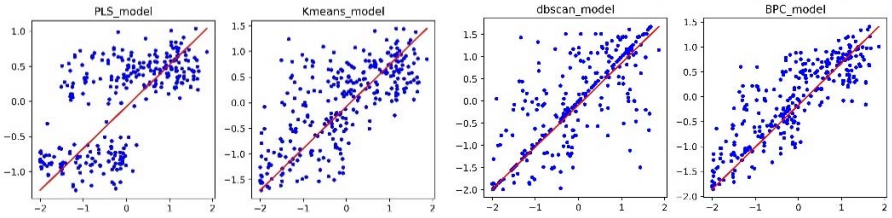

Fig.2 The effect of different methods (The vertical and abscissa represent true and predicted values, respectively)

To explain the advantages of the proposed method, the single-model PLS method and the multimodal method corresponding to other different clustering methods are used for comparison. The results are shown in Fig. 3. The data distribution here is approximately close to the line, indicating that the effect is more accurate. The errors of different methods are expressed by mean absolute error (MAE) and mean square error(MSE), and all the results are shown in Table 2.

Table 2. THE MAE AND MSE OF DIFFERENT METHODS

Methods	MAE	MSE
PLS	0.664	0.630
Kmeans_based	0.5835	0.561
DBSCAN_based	0.453	0.518
BPC_based	0.448	0.341

V. Conclusion

In this paper, a multi-mode soft sensor method based on BPC is proposed. The proposed method aggregates the process data into different sub-blocks to achieve a detailed description of each mode. Combined with PLS, the proposed method can accurately predict the output variables. Experimental results of three-phase flow show that the accuracy of the proposed method is 17.6% higher than that of other methods (measured by MSE). However, since clustering methods are involved, selecting the appropriate number of clustering is still a problem worth discussing. Further analysis of this problem can be conducted in the future.

References

[1] De Assis, Adilson José, and Rubens Maciel Filho. "Soft sensors development for on-line bioreactor state estimation." Computers & Chemical Engineering 24.2-7 (2000): 1099-1103.

[2] Kadlec, Petr, Bogdan Gabrys, and Sibylle Strandt. "Data-driven soft sensors in the process industry." Computers & chemical engineering 33.4 (2009): 795-814.

[3] X. Chen and C. Zhao, "Multivariate Time Delay Estimation Based on Dynamic Characteristic Analytics," 2020 39th Chinese Control Conference (CCC), Shenyang, China, 2020, pp. 2306-2311.

[4] Facco, Pierantonio, et al. "Moving average PLS soft sensor for online product quality estimation in an industrial batch polymerization process." Journal of Process Control 19.3 (2009): 520-529.

[5] Liu, Guohai, et al. "Model optimization of SVM for a fermentation soft sensor." Expert Systems with Applications 37.4 (2010): 2708-2713.

[6] Quiñones-Grueiro, Marcos, et al. "Data-driven monitoring of multimode continuous processes: A review." Chemometrics and Intelligent Laboratory Systems 189 (2019): 56-71.

[7] Ma, Hehe, Yi Hu, and Hongbo Shi. "A novel local neighborhood standardization strategy and its application in fault detection of multimode processes." Chemometrics and Intelligent Laboratory Systems 118 (2012): 287-300.

[8] Yuan, Xiaofeng, et al. "Soft sensor for multiphase and multimode processes based on gaussian mixture regression." IFAC Proceedings Volumes 47.3 (2014): 1067-1072.

[9] Li, Jinna, et al. "Adaptive fault detection for complex dynamic processes based on JIT updated data set." Journal of Applied Mathematics 2012 (2012).

[10] Averbuch-Elor, Hadar, Nadav Bar, and Daniel Cohen-Or. "Border-Peeling Clustering." IEEE transactions on pattern analysis and machine intelligence 42.7 (2019): 1791-1797.