A multimodal soft sensor method for industrial processes based on manifold learning

Guohong Qin[1], Zilu Zhu[2], Qingxi Zheng[3]

[1]School of Iflytek Artificial Intelligence, Chongqing Finance and Economics College, Chongqin, 401320

[2]Ming de College, Northwestern Polytechnical University, Xi an, Shaan xi, 710000

[3]School of Physics and Physical Engineering, Qufu Normal University, Jining, Shandong, 273165

Keywords: multimode, soft sensor, industrial process, Manifold learning, Partial least squares.

Abstract: Data-driven soft sensor technology is essential for modern industrial processes. However, traditional soft sensor techniques usually assume that the data is in a single mode. Due to the switching of working conditions, industrial data often presents multimodal characteristics, which cannot be covered by the single model method. In this work, a multimodal partial least squares method based on manifold learning is proposed to measure industrial processes' key variables. First, considering that process variables' dimension is higher, and the complex process data is distributed in a manifold space, the original information is reduced to a smaller size by t-distributed stochastic neighbor embedding (t-SNE). Then the data belonging to different modes are clustered. Finally, in each mode, the partial least squares model is established to obtain measured variables' values. A real industrial case demonstrates the effectiveness and superiority of the algorithm.

1. Introduction

In the actual industrial production process, due to the complexity of the production process, the measuring environment's degradation, the expensive measuring instruments, and the lagging of the measuring time, the real-time monitoring of product quality appear very difficult [1]-[3]. With the development of artificial intelligence and database technology, data-driven soft sensor technology has attracted extensive attention. Unlike the soft sensor technology based on mechanism analysis, the data-driven method has a low requirement for prior knowledge, and the modeling cost is also lower. It can even consider the impact of noise in the data acquisition process, thus significantly increasing the method's robustness, which is lacking in the traditional modeling method.

The essence of the soft sensor is a regression problem. The simplest regression method is the least-squares (LS) [8] method, which establishes the relationship between multiple measurement variables X and one test variable y. The least-square has a straightforward form and can be easily solved analytically. However, LS has the problem of matrix inversion. This method will fail when collinearity exists between variables X. Therefore, some scholars have added the identity matrix to the correlation matrix to solve the collinearity problem is called ridge regression (RR) [9]. In an industrial process, not every variable collected is useful for the soft sensor. The introduction of too many variables sometimes brings about noise and even reduces the soft sensor model precision. Therefore, modeling methods based on latent variables are widely studied. Principal component regression (PCR) is a typical latent variable modeling method. It first uses principal component analysis (PCA) to extract the latent variables with large variance, representing most of the data's fluctuation information. The latent variables with the most significant fluctuation are selected to establish a relationship with the test variables. Recently, slow feature analysis (SFA) has attracted many scholars [10], which extracts the latent variables that change slowly in the variables for soft measurement. Generally speaking, the latent variables that change slowly can be considered the essential information in the process, while the latent variable that changes quickly can be regarded as noise.

However, in the actual industrial process, there is often more than one variable to be tested. The methods above, including LS, PCR, and SFA, only establish the relationship between multiple input variables and one output variable but ignore the correlation between multiple output variables. To solve this problem, partial least squares (PLS) [3] are proposed and widely explored. Partial least squares regression reduces predictive variables to a smaller set of unrelated components and performs least squares regression on those components rather than on the original data. PLS regression is instrumental when the predictive variables are highly collinear or when the predictive variables are more than the observed values, and the standard error in coefficients produced by ordinary least square regression is high. Unlike multiple regressions, PLS does not assume that the predictive variables are fixed, which means that the measurement of predictive variables may have errors, making the PLS measurement more uncertain.

However, most traditional modeling methods assume that the data is distributed in a mode. Therefore, their application in industrial production is minimal. Due to the switching of working conditions, process transformation, or different loads, the correlation between process variables is not to maintain the original model but to change in a large or small way. This problem, known as the multimode soft sensor, cannot be solved using the methods mentioned above. Cheng et al. applied just-in-time learning (JITL) to deal with the multimode problem [4]. Wang et al. classified the models from a probability distribution, thus realizing a multimodal soft sensor [5]. A k nearest neighbor (KNN) based Fault detection and diagnosis method is developed by Song et al. for the multimode process [6]. However, KNN only uses the Euclidian distance between samples to measure the similarity of models. In some approaches, the higher-dimensional data may be manifold distributed. In other words, Euclidean proximity does not mean that the samples are similar. According to the concept of manifold learning, higher-dimensional data can be embedded in lower-dimensional Spaces. t-Distributed Stochastic Neighbor Embedding (t-SNE) [7] is a typical manifold learning method. Manifold learning can be used to embed high-dimensional data into low-dimensional space, thus mining deeper relationships between variables.

In this paper, a multimode soft sensor method based on manifold learning and PLS is proposed. First, t-SNE is used to project high-dimensional data into low-dimensional space to mine potential connections between samples. Secondly, KNN is applied to search for specimens of different clusters, and the samples are divided into different modes. Finally, PLS is developed to model each mode and establish the relationship between the process data and the output variables.

The rest of the article is arranged as follows. t-SNE and PLS are introduced in section 2. The methods proposed are described in detail in section 3. In section 4, we verify the validity of the algorithm through a three-phase flow experiment. In the last section, some conclusions are drawn.

2. Related work

2.1 t-SNE

t-SNE is a nonlinear manifold dimensionality reduction algorithm for high-dimensional data. This method converts the distance to probability distribution by using Gaussian distribution in high-dimensional space and t-distribution in low-dimensional space to avoid the crowding problem caused by SNE.

In the SNE algorithm, $p_{j|i}$ and $q_{j|i}$ represent conditional probability, x_i 、 x_j 、 y_i and y_j illustrate the similarity adjacent data points, and the conditional probability $p_{j|i}$ conforms to the Gaussian probability distribution.

$$p_{j|i} = \frac{\exp(-\| x_i - x_j \|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\| x_i - x_k \|^2 / 2\sigma_i^2)} \tag{1}$$

Where $\|x_i - x_j\|^2$ represents the square of the distance between x_i and x_j of the adjacent point and σ_i is the variance of a Gaussian function centered on the data point x_i. To obtain the optimal simulation point in the low-dimensional space, the K-L distance between $p_{j|i}$ and $q_{j|i}$ should be minimized.

Cost function C is expressed as

$$C = \mathrm{KL}(P \| Q) = \sum_i \sum_j p_{j|i}^{\log\frac{p_{j|i}}{q_{j|i}}} \tag{2}$$

Where y_i is the minimum gradient descent value in a low-dimensional space.

t-SNE replaces the high-dimensional spatial data conditional probability distribution change to the joint probability distribution. $p_{j|i}$ represents the high-dimensional data points and $q_{j|i}$ illustrates the low-dimensional spatial data points. We define p_{ij} and q_{ij} as

$$\begin{aligned} p_{ij} &= \frac{p_{i|j}+p_{j|i}}{2n} \\ q_{ij} &= \frac{\exp(-\| y_i - y_j \|^2)}{\sum_{k\neq l}\exp(-\| y_k - y_l \|^2)} \end{aligned} \tag{3}$$

Where q_{ij} means the similarity of two points. The new cost function C is expressed as

$$C = \mathrm{KL}(P \| Q) = \sum_i \sum_j p_{ij}^{\log\frac{p_{ij}}{q_{ij}}} \tag{4}$$

Which is equivalent to the joint probability distribution P and Q.

2.2 Partial least squares

PLS is a model based on Global forecast, which constructs the most extensive covariance matrix between the latent variables in the process variable X and the quality data. Thus, the mapping of process variable X to response variable Y can be realized. Here, X is the process variable, Y is the quality variable, n is the number of samples, P is the number of process variables, p is the number of process variables, and L is the number of quality variables. PLS contains the relationship between two types of data, including external relationship and internal relationship.
The external relationship is expressed as

$$\begin{aligned} X &= TP^T + E \\ Y &= TQ^T + F \\ T &= XW(P^TW)^{-1} \end{aligned} \tag{5}$$

Where T, P, and E are the score, load, and residual matrix of X, respectively, Q and F respectively represent the loading matrix and the residual matrix of Y, and W is the weight coefficient matrix.
It also requires an internal relationship between T and U, which is written as

$$\begin{aligned} u_a &= b_a t_a \\ \beta_{pls} &= W(P^TW)^{-1}Q^T \end{aligned} \tag{6}$$

Where Y is the predicted value of X, and B is the regression coefficient

3. Method

3.1 Motivation

Due to the complexity of the production process and the harsh measurement environment, the process data has high dimension and redundancy characteristics. To accurately monitor the working process under different conditions, the process data is generally divided into different modes.

Therefore, the multimode strategy has been widely used in industrial process monitoring. The multimodal problem's research emphasis is how to classify the data with different characteristics and establish the model reasonably. Traditional clustering methods, such as K-means, only consider the Euclidean distance between data. Due to the complexity of the process data, when the data is distributed in a manifold space, Euclidean distance cannot describe the data's length correctly. In the description of The Euclide distance, each dimension's effect on the result is the same, without considering the intrinsic structural characteristics between the sample data. Manifold learning has attracted extensive attention in recent years. By manifold learning, high-dimensional data can be mapped to low-dimensional, and the local structure among the data is preserved, which explores the characteristics among the data. Besides, the influence of process data noise is deleted, and a more effective partition effect can be obtained, which improves the modeling accuracy.

3.2 The proposed model

Due to the switching of working conditions and the change of products in the actual industry, the industrial process often presents typical multimodal characteristics. Besides, the relationship between variables is not only linear but also nonlinear. Considering the data's multimodal and nonlinear features, it is difficult to obtain more accurate results using traditional linear methods. Therefore, it is reasonable to divide the data into different modes and build a model for each mode. Since local data variables' properties are similar, the simple linear model can be applied to analyze them. A multimodal soft sensor method is proposed in this work, which combines t-SNE and partial least squares (PLS). The detailed steps are given as follows.

Step 1: Dimension reduction. Standard dimensional reduction algorithms do not consider the manifold structure between data, such as PCA. However, from the perspective of manifold learning, the observed data is mapped from a low-dimensional manifold to a high-dimensional space. Due to the limitation of the data's internal characteristics, some data in higher dimensions will generate redundancy on the dimension. In fact, only the smaller size can be uniquely represented. Therefore, in this paper, we use the t-SNE, a typical manifold learning method, to reduce the data's dimension. Given the data set $X=\{x_1,\dots,x_N\}$, where N is the number of samples, the t-SNE is adopted on the data. Finally, the low-dimensional features $H=\{h_1,\dots,h_N\}$ are obtained.

Step 2: Data clustering. Clustering algorithms are applied to these low dimensional features to divided process data into different modes. The data characteristics in each mode are similar, but the data characteristics in other methods are quite different. K-means [11] is used here to cluster the process data. First, k centers are randomly selected for the input feature set $X=\{x_1,\dots,x_N\}$. Secondly, the other data set objects into the collection where the central point is located. Thirdly, calculate the new cluster center. Loop the second and third steps until the end condition is met. Finally, the process data are divided into C clusters. The *k*th collection is denoted as $X_k=\{x_{k1},\dots,x_{kN}\}$, and the corresponding measurements are represented as $Y_k=\{y_{k1},\dots,y_{kN}\}$.

Step 3: Soft sensor modeling. In this step, we establish a PLS model for each model. Since each mode's data characteristics are relatively similar, linear models can be used to analyze the data. PLS considers the coupling relationship between multiple input variables and multiple output variables simultaneously, improving prediction accuracy. And this is a linear model with low computational complexity. For the *k*th mode, the mathematical expression is as follows.

$$\begin{aligned}
&\mathrm{X}_k=\mathrm{T}_k\mathrm{P}_k{}^{\mathrm{T}}+\mathrm{E}_k\\
&\mathrm{Y}_k=\mathrm{T}_k\mathrm{Q}_k{}^{\mathrm{T}}+\mathrm{F}_k\\
&\mathrm{T}_k=\mathrm{X}_k\mathrm{W}_k(\mathrm{P}_k{}^{\mathrm{T}}\mathrm{W}_k)^{-1}\\
&\mathrm{u}_{k,a}=\mathrm{b}_{k,a}\mathrm{t}_{k,a}\\
&\beta_{\mathrm{pls}}=\mathrm{W}_k(\mathrm{P}_k{}^{\mathrm{T}}\mathrm{W}_k)^{-1}\mathrm{Q}_k{}^{\mathrm{T}}
\end{aligned}\tag{7}$$

Step 4: Online application. For online application, when the new sample X_{new} arrives, calculate the distance between the current sample and the data center of each mode to determine the corresponding mode, which is expressed as

$$D_k = \| X_{new} - X_{k-center} \|^2 \tag{8}$$

The corresponding modes are then used to predict the current value.

4. Section 4

4.1 Description of three-phase flow facility

The data in this paper are from the three-phase stream equipment of Cranfield University. The purpose of this equipment is to study the effect of multiphase flow supply on small industrial equipment. The operation process can be roughly described as the oil, water, and gas storage tank of the unit provides oil, water, gas, single-phase or multiphase flowing substances to the equipment. When the reaction is over, they are returned to the tank after being treated by the equipment.

In this article, the sampling interval for all data is 1 second. Three-phase separators are always pressurized to 0.1 MPa. We collected three data sets of operating conditions (T1, T2, and T3), among which the changes of water with air are shown in Table 1.

Table 1. List of process variables used in this study

Air flowrate(m^3/s)	$0.0208m^3/s$	$0.0278m^3/s$	$0.0347m^3/s$	$0.0417m^3/s$
Water flowrate(kg/s)	0.5	1	2	3.5

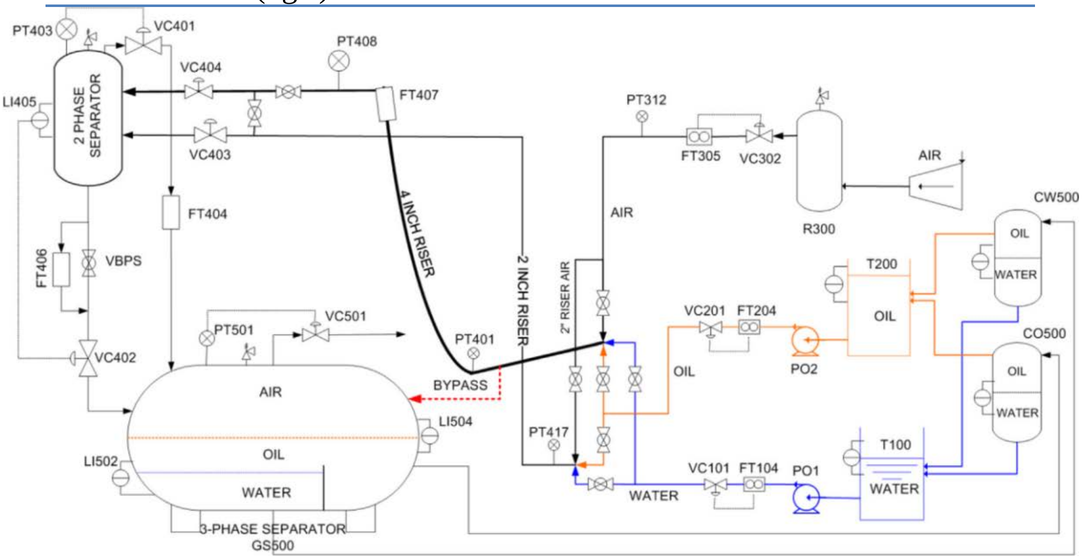

Fig. 1. The sketch of the three-phase flow facility.

The modeling data used in this paper are 19 different input process variables, as shown in Table 2, and the prediction variables are shown in Table 3 with four other variables. In this paper, MAPE is used to measure the relative errors between the average test value and the test set's real value.

Table 2 List of process variables used in this study

Variable nr	Location	Measured magnitude	Unit
1	PT312	Air delivery pressure	MPa
2	PT408	Diff. pressure (PT401-PT408)	MPa
3	PT403	Differential pressure over VC404	MPa
4	FT305	Flow rate input air	Sm3/s
5	FT104	Flow rate input water	kg/s
6	FT407	Flow rate top riser	kg/s
7	LI405	Level top separator	m
8	FT406	Flow rate top separator output	kg/s
9	FT407	Density top riser	Kg/m^3
10	FT406	Density top separator output	Kg/m^3
11	FT104	Density water input	Kg/m^3
12	FT407	Temperature top riser	℃
13	FT406	Temperature top separator output	℃
14	FT104	Temperature water input	℃
15	LI504	Level gas-liquid 3 phase separator	%
16	VC501	Position of valve VC501	%
17	VC302	Position of valve VC302	%
18	VC101	Position of valve VC101	%
19	PO1	Water pump current	A

Table 3. List of process variables predicted in this study

Variable nr	Location	Measured magnitude	Unit
1	PT408	Pressure in top of the riser	MPa
2	PT501	Pressure in 3 phase separator	MPa

4.2 Experiment

In this section, the proposed algorithm's effectiveness and superiority are verified by a three-phase flow process. We first study the internal structural characteristics of the data. However, high-dimensional data is not easy to be presented. To improve the expression ability of data and reduce the training complexity by t-SNE. The data are mapped from the higher dimension to the three dimensions, which is shown in Fig. 2.

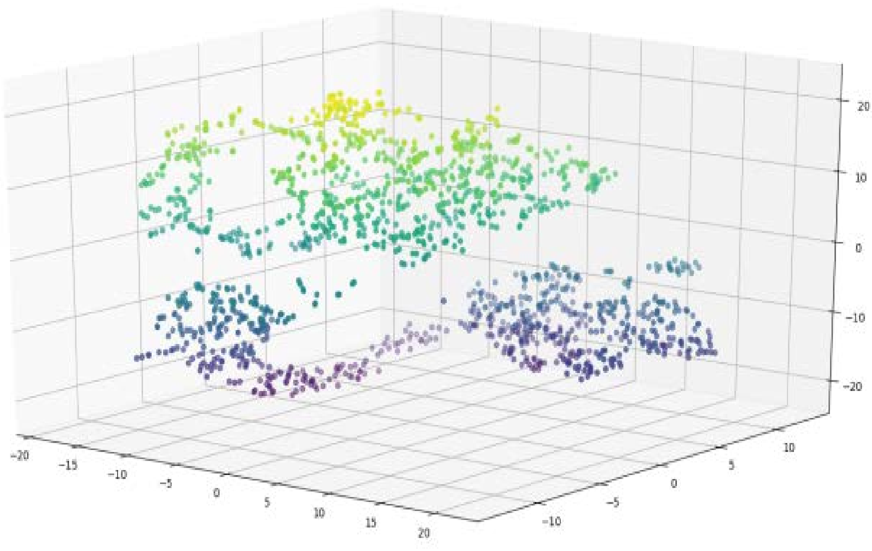

Fig. 2. A low-dimensional structure of data.

To illustrate the advantages of manifold dimensionality reduction, we compare the multimodal method based on PCA (Method 2) and the multimodal algorithm without dimensionality reduction (Method 3). After that, the PLS model is established for each mode, and the values of the three process variables are predicted and analyzed. The results are shown in Table 5.

Table 5. Forecast result

Dimension reduction algorithm	The proposed method	Method 2	Method 3
Pressure in top of the riser	-0.0571	0.8337	-3.2227
Pressure in 3 phase separator	0.0773	0.1646	-14.7879

From the result, we can see that the manifold dimensionality reduction has a better performance, which presents the various structure of data. Besides, Method 2 has a more accurate result than Method 3. It illustrates that dimensionality reduction is meaningful for complex high-dimensional process data. The predictions of the three variables are shown in Fig. 3. The effect of clustering is shown in Fig. 4.

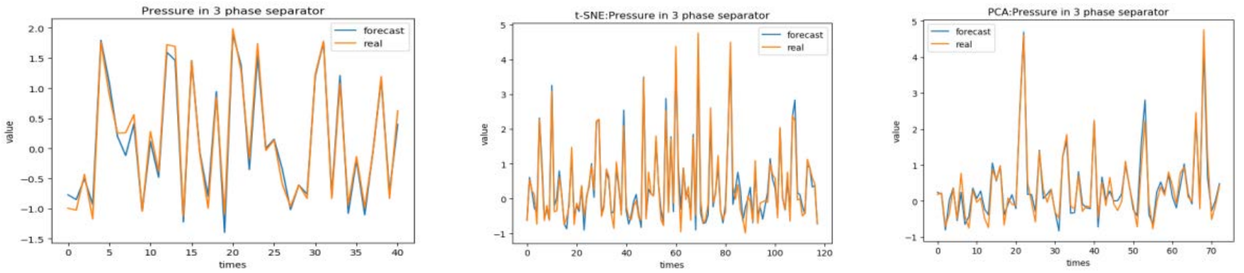

Fig. 3. The predictions of three variables.

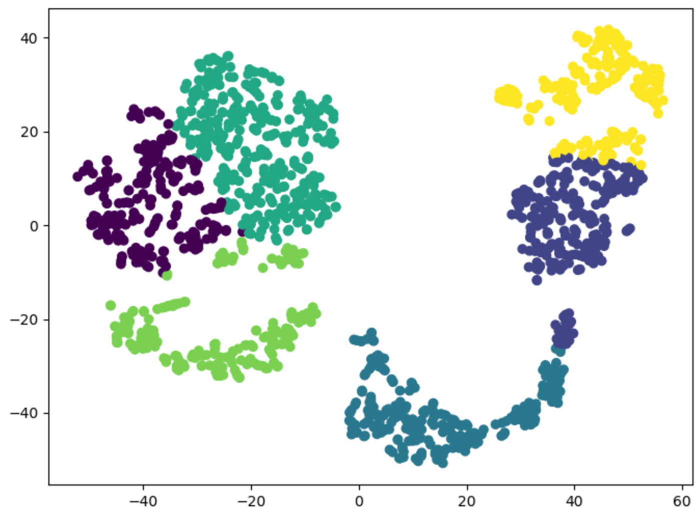

Fig. 4. The result of clustering

5. Conclusion

In this paper, a multimodal partial least squares method based on manifold learning is proposed to measure industrial processes' key variables, which deals with the multimodal characteristics. The manifold structure of process data is reflected by manifold learning, reavealing the essential characteristics between process data. It also provides the basis for mode division and identification. In each mode, a PLS model is established to eatract the latent features and measure the key process variables. However, manifold learning assumes that the data are distributed in the same manifold, which reduces the accuracy of the experiment. In the future work, the multi-manifold learning method can be considered to deal with the multi-modal characteristics to achieve a more accurate pattern division

References

[1] BROSILLO W. Inferential control of process [J].Journal of AIChE, 1978, 24(3):485-509.

[2] HAZAMA K,KANO M. Covariance-based locally weighted partial least squares for high-performance adaptive modeling[J].Chemometrics and Intelligent Laboratory Systems,2015,146:55-62.

[3] MEHMOOD T, LILAND K H, SNIPEN L, et al. A review of variable selection methods in partial least squares regression [J].Chemometrics and Intelligent Laboratory Systems, 2012, 118(16): 62-69.

[4] Cheng Cheng, Min-Sen Chiu, Nonlinear process monitoring using JITL-PCA, Chemometrics and Intelligent Laboratory Systems, Volume 76, Issue 1, 2005, Pages 1-13。

[5] Lulu Wang, Jiusun Zeng, Xiaoyu Liang, Yuchen He, Shihua Luo, and Jinhui Cai, Industrial & Engineering Chemistry Research 2019 58 (31), 14267-14274

[6] Bing Song, Shuai Tan, Hongbo Shi, Bo Zhao, Fault detection and diagnosis via standardized k nearest neighbor for multimode process, Journal of the Taiwan Institute of Chemical Engineers, Volume 106, 2020, Pages 1-8.

[7] van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). "Visualizing Data Using t-SNE". Journal of Machine Learning Research. 9: 2579–2605.

[8] Levenberg, K. (1944). A method for the solution of certain nonlinear problems in least squares. Quarterly of applied mathematics, 2(2), 164-168.

[9] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. Technometrics, 12(1), 69-82.

[10] Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. Neural computation, 14(4), 715-770.

[11] Hamerly, G., & Elkan, C. (2004). Learning the k in k-means. In Advances in neural information processing systems (pp. 281-288).