SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Fault diagnosis of industrial process based on the optimal parametric t-distributed stochastic neighbor embedding

Ruixue Jia, Jing Wang* & Jinglin Zhou*

College of Information Science and Technology
Beijing University of Chemical Technology, Beijing 100029, China

Appendix A Fault classification model based on parametric t-SNE

At present, data-driven fault diagnosis methods for industrial processes are complex and diverse, but the goals are to analyze and process the operational data. The fault diagnosis of the system is done quickly and accurately without knowing the precise analytical model. The industrial process data has many variables, high complexity and strong correlation. The main purpose of adopting the method in fault diagnosis is to extract fault data features through dimensionality reduction, while the extracted fault features maintain the correlation of original data. Finally, fault diagnosis is realized by fault feature classification. Since the parametric t-SNE can preserve local structure and global classification information of fault data in the low-dimensional feature space, the fault classification model based on the parametric t-SNE is established to reveal the nonlinear structure of original data and ensure the accuracy of fault type diagnosis.

In parametric t-SNE, the parametric mapping $f : X \longrightarrow Y$ from the data space X to the low-dimensional feature space Y is parametrized by means of a deep belief network (DBN) with weights W. Then, the network weight is fine-tuned using t-SNE backpropagation as to minimize the cost function that attempts to retain the local structure of the data in the feature space [1]. Therefore, as shown in figure A1, the specific process of establishing offline fault classification model by using t-SNE algorithm is as follows:

(1) Normalize the training sample data $X = \{x_1, \ldots, x_N\}, x_i \in R^D$. Let $V_1 = x_i$.

(2) Perform the first-level RBM training on the preprocessed data. After training, the hidden layer output is used as the input of the second-level RBM, and the next layer of training is continued until each layer was trained successively according to the preset network structure. Then, the network model parameters are obtained.

(3) Fine-tuning: The feature data obtained by inputting the historical fault data into the deep belief network is used as the sample initial solution $y^{(0)}$ in the t-random neighbor embedding algorithm. Then, the network weight is fine-tuned using t-SNE backpropagation as to minimize the cost function. Finally, the optimal model parameters for the global deep belief network are obtained.

Appendix A.1 Deep belief network pre-training

The deep belief network has enough hidden layers (with a nonlinear activation function) to be able to parameterize any complex nonlinear functions. Therefore, a deep belief network is chosen to learn the nonlinear mapping between high-dimensional fault data space and low-dimensional feature space. DBN is composed of multiple layers of RBM, and its training process is: use unsupervised greedy layer by layer method to obtain the weight of pre-training.

Restricted Boltzmann Machine (RBM) is an undirected probability graph model that is interpreted using random neural networks. The structure of RBM is a fully connected bipartite graph which is divided into two layers, the visible layer (input layer) and the hidden layer. The number of neurons in the visible layer is equal to the dimension of the input high-dimensional data, and the number of nodes in the hidden layer is the dimension of the extracted feature vector. The connection between neurons is characterized by no connection within the layer and full connection between layers. Given a training sample, the RBM is trained by adjusting the parameters of RBM model, so that the probability distribution of visible nodes represented by RBM under this parameter is as consistent as possible with the probability distribution of training data.

RBM is a special case of Markov random fields. When a set of visible layer states v and hidden layer states h are given, the joint distribution among all the nodes is represented by the energy function $E(v, h \mid \theta)$,

* Corresponding author (email: jwang@mail.buct.edu.cn, jinglinzhou@mail.buct.edu.cn)

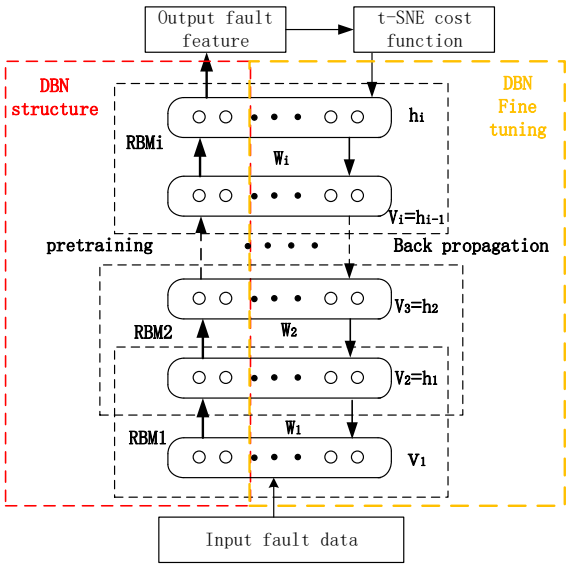

Figure A1 Structure of fault classification model based on parametric t-SNE.

$$E(v,h\mid\theta)=-\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}v_ih_j-\sum_{i=1}^{n}v_ia_i-\sum_{j=1}^{m}h_jb_j, \tag{1}$$

where v_i is the node state of visible layer, h_j is the node state of hidden layer, $\theta=w,a,b$ is the model parameter, w_{ij} is the weight of the connection between nodes v_i and h_j, a_i is the bias on the node v_i, b_j is the bias on the node h_j.

$P(v\mid\theta)$ is obtained by Indexing and regularizing the energy function $E(v,h\mid\theta)$. So $P(v\mid\theta)$ needs to be maximized,

$$P(v\mid\theta)=\sum_{h}\exp^{-E(v,h\mid\theta)}/\sum_{v,h}\exp^{-E(v,h\mid\theta)}, \tag{2}$$

where, $\sum_{v,h}\exp^{-E(v,h\mid\theta)}$ is a normalization factor that represents the sum of all possible states (energy index) of the visible layer and hidden layer node sets; $\sum_{h}\exp^{-E(v,h\mid\theta)}$ represents the probability of the visible layer node set under a certain state distribution. The maximization of the cost function is performed using a gradient ascent method. Model parameter is updated by

$$\theta_{i+1}=\theta_i+\mu\frac{\partial\ln P(v^t)}{\partial\theta_i}, \tag{3}$$

where, μ denotes the learning rate, v^t denotes the input data of the model. Because the joint probability distribution $P(v)$ of the visible layer cannot be obtained directly, we use the contrast divergence CD-k algorithm to reconstruct $P(v)$. Sample data is repeated (4), (5) after iteration k times, the state of the visible layer v_i will converge on the joint probability distribution $P(v_i)$.

$$P(h_j=1\mid v^t)=\frac{1}{1+\exp(-\sum_i v_iw_{ij}-b_j)}, \tag{4}$$

$$P(v_i=1\mid h)=\frac{1}{1+\exp(-\sum_j h_jw_{ij}-a_i)}, \tag{5}$$

Here, because a large amount of the same or similar data is trained in the sample data, It is equivalent to a sample iterating multiple times. Therefore, the number of iterations is taken as $k=1$.

Therefore, the pre-training process of the deep belief network is as follows:

(1) Normalize the training sample data $X=\{x_1,\ldots,x_N\},x_i\in R^D$. Let $V_1=x_i$.

(2) Perform the first-level RBM training on the preprocessed data. After training, the hidden layer output is used as the input of the second-level RBM, and the next layer of training is continued until each layer was trained successively according to the preset network structure.

Appendix A.2 t-SNE backpropagation fine-tuning model parameters

The t-Distributed Stochastic Neighbor Embedding (t-SNE) [1] is a machine learning algorithm for dimensionality reduction. In addition, t-SNE is a non-linear dimensionality reduction algorithm, which is very suitable for visualizing high-dimensional data by dimension reduction to 2D or 3D. Compared with other dimensionality-reduction techniques, t-SNE can can preserve the local structure and global structure in low-dimensional space. In other words, it has a certain degree of clustering effect and better visualization of high-dimensional data to reduce dimension to a 2-dimensional plane.

The t-SNE starts by converting the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities. Specifically, given a training sample set $\{x_1, \ldots, x_N\}, x_i \in R^d$, the conditional probability of data point x_j to data point x_i is $p_{j|i}$, and $p_{j|i}$ is proportional to similarity between x_i and x_j. The conditional probability $p_{j|i}$ is given by

$$p_{j|i} = \frac{\exp(-\parallel x_i - x_j \parallel^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\parallel x_i - x_k \parallel^2 / 2\sigma_i^2)}, \tag{6}$$

where, σ_i is the Gaussian variance centered on data point x_i. For different points x_i, σ_i does not have the same value. It is not possible for σ_i to be optimal for all data points in the data set because the density of the data may change. Therefore, the t-SNE performs the value of the binary search σ_i using the fixed confusion level $Perp$ specified by the user. The usual perplexity level is between $5 - 50$. The relationship between perplexity $Perp$ and σ_i is as shown in equations (7) and (8)

$$Prep(P_i) = 2^{H(P_i)}, \tag{7}$$

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}, \tag{8}$$

In t-SNE, the degree of confusion is the parameter we set primarily, which is roughly equivalent to the nearest neighbors that are considered when matching the original and fitted distribution of each point. For different data sets, we need to adjust the confusion by experiment. The specific setting of the degree of confusion $Perp$ performs the binary search σ_i value as follows:

1) Calculate the Euclidean distance square between the data points from the sample set X, and construct a pairwise distance matrix D, where the $i-th$ row and the $j-th$ column element are the Euclidean distance square between the input points x_i and x_j.

2) Convert the pairwise distance matrix D between data points into a conditional probability matrix P, where the $i-th$ row and the $j-th$ column elements are $p_{j|i}$. At this time, it is necessary to perform a binary search by setting the confusion degree $Perp$ to set a σ_i for each line of the matrix P.

3)The upper limit value and the lower limit value of σ_i are set. The tolerance of $Perp$'s tolerance is set. Then, Each time the center value of the σ_i range is taken, the confusion degree $Perp$ is calculated by substituting equations (6), (8), and (7) until the searched σ_i satisfied the $Perp$ tolerance condition, and the search is stopped.

Also set $p_{i|i} = 0$, because we are concerned about the similarity between the two. In order to optimize the subsequent cost function, the conditional probability distribution is replaced with the joint probability distribution. In this case, let $p_{ij} = (p_{j|i} + p_{i|j})/2n$.

For the low dimensional corresponding points y_i and y_j of the high dimensional data points x_i and x_j, we use q_{ij} to represent a similar joint probability. In order to solve the crowding problem of reducing the dimensionality of high-dimensional data points into low-dimensional space, in a low-dimensional map, a one-degree-of-freedom t-distribution is used instead of a Gaussian distribution to express the similarity between two points. The joint probability q_{ij} is defined as

$$q_{ij} = \frac{(1+\parallel y_i - y_j \parallel^2)^{-1}}{\sum_{k \neq l}(1+\parallel y_k - y_l \parallel^2)^{-1}}, \tag{9}$$

We model the similarity between the low-dimensional data points y_j and y_i: The overall goal is to select one of the data points in Y, and then make its joint probability distribution q approximate to p. Therefore, the Kullback Leibler divergence is used to measure the probability distribution Q to fit the true distribution P. Objective function C is defined by

$$C = KL(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{10}$$

t-SNE uses the gradient descent method to minimize the objective function C. The complete gradient formula after derivation is defined by

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1+\parallel y_i - y_j \parallel^2)^{-1}, \tag{11}$$

The gradient $\frac{\partial C}{\partial y}$ is calculated by Equation (11), and the low-dimensional data is updated by

$$y^{(t)} = y^{(t-1)} + \eta \frac{\partial C}{\partial y} + \alpha(t)(y^{(t-1)} - y^{(t-2)}), \tag{12}$$

Here, η is learning rate, $\alpha(t)$ is momentum, $y^{(t)}$ is the set of low-dimensional data after t times update. The feature data obtained by inputting the historical fault data into the deep belief network is used as the sample initial solution $y^{(0)}$. In this way, the initial low-dimensional data is located in a better position, which can accelerate the target value to converge to a better local extreme value in the process of gradient descent. The target result is the low dimensional data $y^{(T)}$ after iterating T times.

Therefore, denoting the mapping from the data space to the feature space that is defined by the DBN's pre-training network as $f : X \longrightarrow Y$, this leads to the following definition.

$$q_{ij} = \frac{(1+\parallel f(x_i \mid W) - f(x_j \parallel W) \parallel^2)^{-1}}{\sum_{k \neq l}(1+\parallel f(x_k \mid W) - f(x_l \parallel W) \parallel^2)^{-1}}, \tag{13}$$

The minimization of the cost function C (defined by equation (13) using q_{ij}) can be performed using backpropagation, and the gradient required for fine tuning is given by

$$\frac{\partial C}{\partial W} = \frac{\partial C}{\partial f(x_i \mid W)} \frac{\partial f(x_i \mid W)}{\partial W}, \tag{14}$$

where, $\frac{\partial f(x_i \mid W)}{\partial W}$ is calculated using standard backpropagation, $\frac{\partial C}{\partial f(x_i \mid W)}$ is calculated as follows:

$$\frac{\partial C}{\partial f(x_i \mid W)} = 4 \sum_j (p_{ij} - q_{ij})(f(x_i \mid W) - f(x_j \mid W))(1 + \| f(x_i \mid W) - f(x_j \mid W) \|^2)^{-1}, \tag{15}$$

Therefore, the specific steps for feature extraction based on the parametric t-SNE are:

(1) The fault data is input into deep belief network for pretraining and the model parameters are obtained.

(2) The feature data obtained by inputting the historical fault data into the deep belief network is used as the sample initial solution $y^{(0)}$ in the t-random neighbor embedding algorithm. Then, the network weight is fine-tuned using t-SNE backpropagation as to minimize the cost function. Finally, the optimal model parameters for the global deep belief network are obtained. Meanwhile, the local structure of the original data is preserved in the low-dimensional map.

Appendix A.3 Selection of the optimal parameter value for the parametric t-SNE algorithm

The parametric t-SNE algorithm has only one parameter "perplexity" that must be pre-specified by the user, but no guidance was yet given how to choose it. Therefore, a parameter optimization indicator is defined to measure the "quality" of the input-output mapping, and automatically select the optimal parameter value based on the indicator. We perform experimental verification of industrial fault classification under the optimal parameter values and get the best classification effect.

The perplexity can be interpreted as a smooth measure of the effective number of neighbors, and typical values are between 5 and 50. The value of perplexity is proportional to the number of nearest neighbors. A large number of nearest neighbors causes smoothing or eliminating of small-scale structures in the manifold. In contrast, too small neighborhoods can falsely divide the continuous manifold into disjoint sub-manifolds [3]. Therefore, we need to select the best value of perplexity according to the indicators defined below, so that the classification model has the best classification accuracy.

We use kullback leibler divergence to measure the "quality" of the input-output map, that is, the degree to which the high-dimensional structure is represented in the embedded space. Evaluation function D is defined as

$$D = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{16}$$

p_{ij} and q_{ij} are the probability distribution of data in X and Y, respectively. The lower the value of D is, the better high-dimensional data are represented in the embedded space. Hence, the optimal value for $Prep$, $Prep_{opt}$, can be determined as

$$Prep_{opt} = \arg \min_{Prep} D \tag{17}$$

Here, a method to determine $Prep_{opt}$ is to run parameter t-SNE with every possible $Prep$ ($Prep \in [5, 50]$) and select $Prep_{opt}$ according to Eq.17.

Appendix A.4 Online implementation of fault classification model

The K nearest neighbor algorithm is a classification algorithm. The KNN predicts the type of new samples by calculating the distance between the new sample data and the nearest neighbor k historical sample data points. In this paper, firstly, the high-dimensional fault data of industrial process is projected to the optimal classification feature space by using the parametric t-SNE. Then, in the feature space, the K-nearest neighbor algorithm is used to predict the type of new fault data online. Specific steps are as follows:

(1)Real-time fault data of industrial process is projected to the optimal classification feature space by using the parametric t-SNE.

(2) In the feature space, calculate the Euclidean distance between real-time data and each known types of training data.

(3) European distances are Sort in ascending order.

(4) Select the K points with the smallest distance and calculate the frequency of occurrence of the category of the K points.

(5) Return the category with the highest frequency among the top K points as the prediction category of the real-time data.

Appendix B Experimental verification and analysis

Appendix B.1 MNIST data set experimental verification

In this paper, the MNIST data set, including 70,000 handwritten-number grayscale images, is selected for experimentation. Each image consists of $28 \times 28 = 784$ pixels, and each pixel is represented by a grayscale value. In this experiment, each picture is transformed into a sample of a row vector in 1×784 dimension. The training set includes 30,000 samples

containing handwritten numbers $1 \sim 5$ and the test set includes 5,000 samples. First, we use the parameter optimization index proposed in the paper to select the optimal *Prep* value of 30 for the parametric t-SNE. Figure B1 (a) and (b) show the visualization of the high-dimensional data of 5,000 test samples which is reduced to the 2-dimension plane using FDA and parametric t-SNE methods, respectively. Parametric t-SNE shows good clustering and visualization effects compared with the traditional FDA method. What is the influence of the feature number on the classification performance? Here, we map the MNIST dataset into different feature spaces in order to answer this question. Figure B2 (a)-(c) shows the classification recognition rate of test set at different the reduction dimension space with d=2, 10 and 30, respectively. It is found that the parametric t-SNE always shows better recognition rate in different feature space than FDA method. It has obvious advantages in the 2-dimensional space, especially. It also is found from Figure B2 (d) that the average recognition rates of all categories using the two methods are improved with the increase of the feature number. Table B1 gives the average recognition rates of test data. It is found from Table B1 that the parametric t-SNE can can accurately cluster data by extracting fewer features. The parametric t-SNE displays better classification because it retains the local structure of original data and global classification information after the dimension reduction.

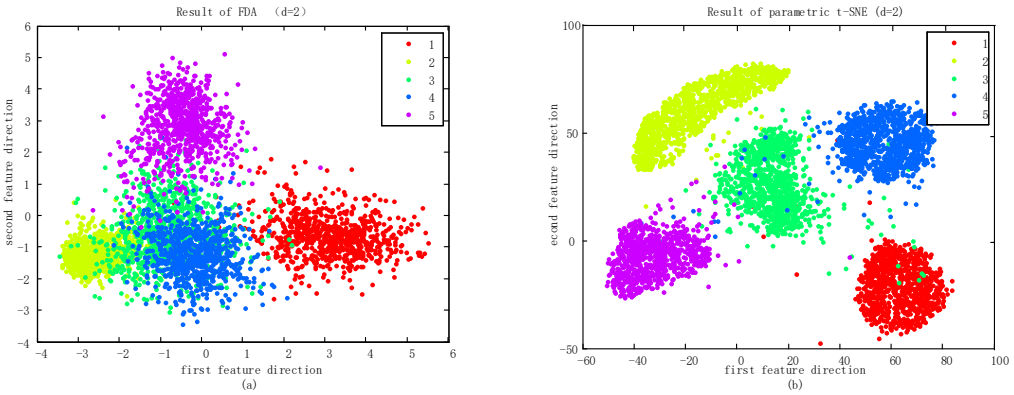

Figure B1 Visualization of 5,000 handwritten digits of MNIST data: (a) FDA; (b) parametric t-SNE.

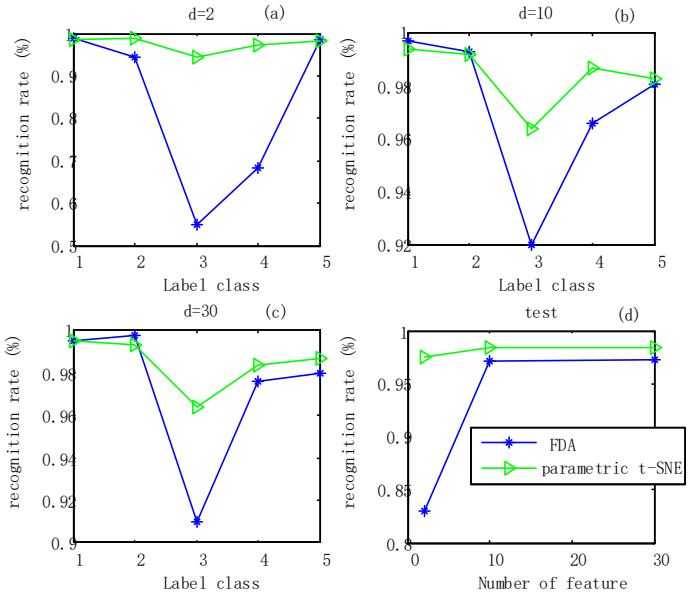

Figure B2 The effect of reduction dimension on classification: (a) d=2, (b) d=10, (c) d=30.

Appendix B.2 TE process fault classification verification

The Tennessee Eastman Process (TEP) is a simulation platform of an actual chemical process which is a benchmark for testing the effectiveness of fault diagnosis or process control methods. TEP includes five major units: a chemical reactor,

Table B1 Classification comparison of FDA and parametric t-SNE on MNIST data set

	d=2	d=10	d=30
FDA	0.830	0.971	0.972
parametric t-SNE	0.975	0.984	0.984

a condenser, a recycle compressor, a vapor/liquid separator, and a stripper. Figure 5 shows a detailed flow chart of TEP. This process provides 52 observation variables, consists of 41 process variables and other 11 manipulated variables. TEP is a typical non-linear, strongly coupled and dynamic system. 21 pre-defined faults are given in TEP including 16 known faults and 5 unknown faults. The specific fault types are shown in Table B2. Here, Faults 1 to 7 are related to step changes. Faults 8 to 12 are caused by random variables of the process. Fault 13 is a slow offset. Faults 14, 15 and 21 are caused by valve sticking. Faults 16 to 20 are unknown faults. It is pointed that Faults 1, 2, 6, 7, 8, 12, 13, 14, 17 and 18 are significant faults, and faults 3, 4, 5, 9, 10, 11, 15, 16, 19, 20 and 21 are minor faults [4]. There are 22 data sets in total including the normal and different fault operations. The data consists of 22 different simulation run data. The entire operation take 48 hours and the fault is introduced at the 8th hour. The data collection time is 3 minutes. The normal operation data contains 960 samples, and each of the 21 different fault training data contains 800 samples.

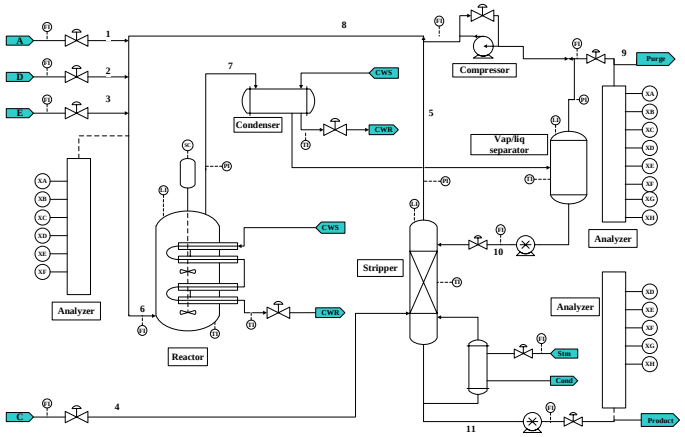

Figure B3 TE process flow chart.

Table B2 Faults in Tennessee Eastman process

No.	Description	Type
IDV(0)	Normal operation	–
IDV(1)	A/C feed ratio, B composition constant	Step
IDV(2)	B composition, A/C ratio constant	Step
IDV(3)	D feed composition temperature	Step
IDV(4)	Reactor cooling water temperature	Step
IDV(5)	Condenser cooling water inlet temperature	Step
IDV(6)	A feed loss	Step
IDV(7)	C header pressure loss-reduced availability	Step
IDV(8)	A, B, C feed composition	Random variation
IDV(9)	D feed composition temperature	Random variation
IDV(10)	C feed composition temperature	Random variation
IDV(11)	Reactor cooling water temperature	Random variation
IDV(12)	Condenser cooling water inlet temperature	Random variation
IDV(13)	Reaction kinetics	Slow drift
IDV(14)	Reactor cooling water valve	Sticking
IDV(15)	Condenser cooling water valve	Sticking
IDV(16)	Unknown	Unknown
IDV(17)	Unknown	Unknown
IDV(18)	Unknown	Unknown
IDV(19)	Unknown	Unknown
IDV(20)	Unknown	Unknown
IDV(21)	The valve fixed at steady state position	Unknown constant position

Here, fault 2, fault 4 and fault 6 are selected in order to evaluate the fault classification performance of the parametric t-SNE method. These three faults are caused by step changing of different operation condition and fault 4 is a minor fault which is difficult to be detected. Because the actual data collected from the TEP is limited, the training model needs enough training data. Therefore, the operation of each type of fault are run 10 times. Then, models are built using 6000 training data points and the 2000 test data points are used to test the performance of the model. We need to select different optimal parameter $Prep$ to achieve the optimal classification performance for different data sets. Figure B4 (a) shows a plot of the parameter optimization evaluation function D for $Prep$ ranging from 5 to 60. It can be found that the value of this function has a minimum value with the change of $Prep$. When the value of $Prep$ is 35, the value of kullback leibler divergence is the smallest. In order to verify the effectiveness of the parameter optimization method, figure B4 (b) shows

the relationship between the fault type recognition rate of test data and the value of *Prep*. It can be found that the fault classification performance of the parametric t-SNE is optimal when parameter *Prep* is selected as 35. Therefore, the *Prep* is determined to be 35 under the premise of ensuring a high fault recognition rate.

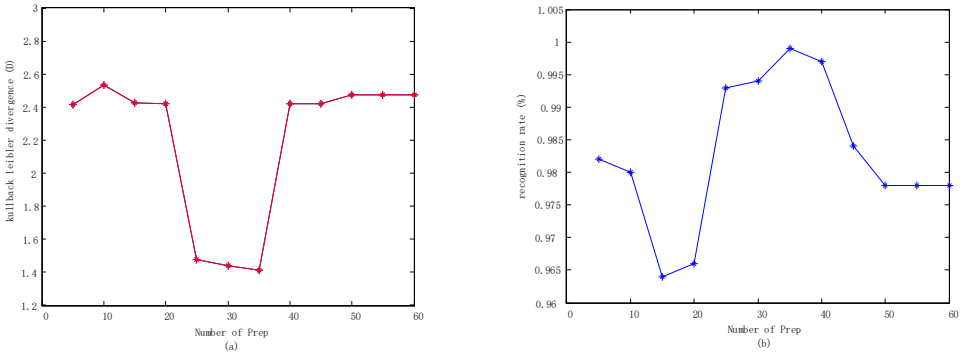

Figure B4 (a)Selection and (b)verification of the optimal parameter value for the parameter t-SNE algorithm

To visually show the characteristics of the different fault classification and diagnosis methods, we extracted 2 features and plot the fault data projection results in Figure B5. The visual effects of the parametric t-SNE on the classification of three faults are compared with that of FDA and LLEDA methods. The parametric t-SNE has better classification effects on fault data than FDA and LLEDA, and it has less uncertainty. Table B3 shows the correct recognition rate for three types of faults based on LLEDA, FDA, and parametric t-SNE methods, respectively. From the data in the table, the parametric t-SNE has a good accuracy rate for fault classification diagnosis. Moreover, because the t-SNE and LLEDA methods preserve local geometry and global information in the process of nonlinear industrial data dimensionality reduction, the FDA linear dimensionality reduction process may not reveal the nonlinear structure contained in the data set. The fault recognition rate of the t-SNE and LLEDA methods is significantly better than the FDA method. At the same time, the parametric t-SNE and LLEDA have similar judgment performance. However, it is found from the experimental results that as the number of features of the fault data increases, the fault recognition rate of the parametric t-SNE will be slightly better than LLEDA. This is because the actual industrial fault data is uncertain and non-linear. Although the LLEDA algorithm can also implement nonlinear data classification, its reconstruction weight is a linear Gaussian assumption, which sometimes has no advantage for non-Gaussian data. However, the parametric t-SNE solves the problem of data distribution by converting the data distance problem into a probability distribution problem. Therefore,the validity of the parametric t-SNE for industrial process fault classification diagnosis is verified by experimental data, and the parametric t-SNE has better clustering effect on the extracted features of fault data than FDA and LLEDA. Moreover, the parametric t-SNE can better distinguish non-Gaussian nonlinear industrial fault data without more features.

Table B3 The correct recognition of faults 2, 4, and 6 by different methods

Number of features	Recognition rate	FDA	LLEDA	parametric t-SNE
2	Fault 2	0.965	1	1
	Fault 4	1	1	1
	Fault 6	1	1	1
3	Fault 2	0.965	1	1
	Fault 4	1	1	1
	Fault 6	1	1	1
4	Fault 2	0.965	0.993	1
	Fault 4	1	1	1
	Fault 6	1	1	1
5	Fault 2	0.963	0.990	1
	Fault 4	1	1	1
	Fault 6	1	1	1

Appendix B.3 Penicillin Fermentation Process classification verification

Penicillin is a widely used antibiotic in humans, and its fermentation process is a complex nonlinear batch process. Therefore, the research on the monitoring of this process has great practical significance. The penicillin fermentation reaction process is characterized by continuity, nonlinearity and uncertainty. Figure 8 shows the basic flow chart of the process. In this paper, we use the simulation platform Pensim 2.0 software of penicillin fermentation process to carry out simulation experiments. Pensim 2.0 was developed in $1998-2002$ by the Institute of Process Modeling, Monitoring and Control, led by Prof. Cinar, by the Illinois Institute of Technology. This simulation platform is specially designed for the penicillin fermentation process, which can realize a series of simulations of penicillin fermentation process. Relevant research has shown the practicability

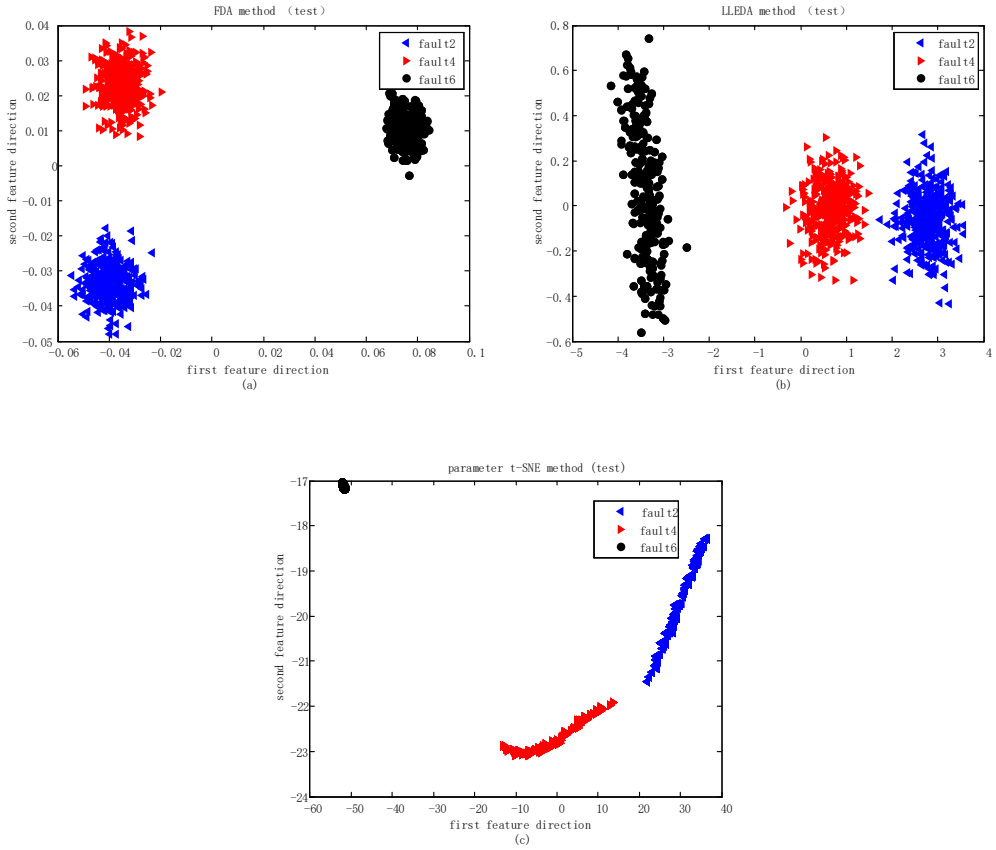

Figure B5 Classification visualization of 1000 failure data for fault 2,4,6: (a) FDA; (b) LLEDA; (c) parametric t-SNE.

and effectiveness of the simulation platform. Therefore, Pensim 2.0 has become an effective way for many scholars to study the diagnosis and monitoring of intermittent process faults [5].

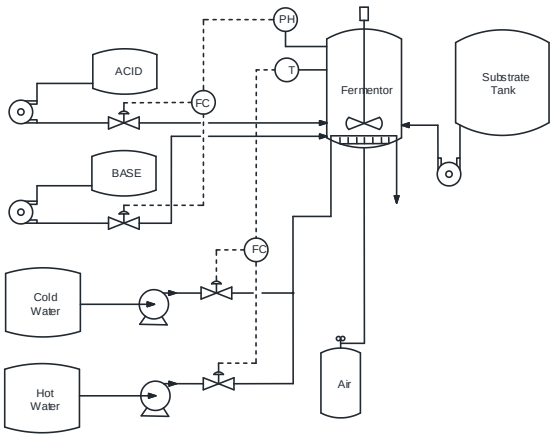

Figure B6 Flow diagram of the penicillin fermentation process.

In this paper, we use the simulation platform Pensim 2.0 of the penicillin fermentation process to collect data. Because the variables such as substrate concentration and microbial concentration cannot be measured in real time, the data of some variables is defective. Here, 9 variables are selected to built model of the fault classification (as shown in Table B4). Three types of simulated faults (as shown in Table B5) are chosen to verify parametric t-SNE method. Firstly, the 30 batches of data obtained by Pensim 2.0 simulation are used as the training data set, the reaction time is 400 hours, the sampling time is 0.5 hours, including 10 batches of fault data of air flow drop, 10 batches of fault data of stirring power drop and

10 batches of fault data of bottom materials feed rate drops. Each batch of fault data is obtained by changing the size of the fault. Here, we establish a fault classification model of the proposed method by using each type of 4000 fault data as a training sample, and then project the test fault data into the two-dimensional map through the model to visualize. In order to get the optimal fault classification model, we use the parameter optimization index proposed in the paper to select the optimal confusion value of 35.

Table B4 Variables used in the monitoring of the benchmark model

No.	Variable
1	Agitator power
2	Aeration rate
3	Substrate feed rate
4	Substrate feed concentration
5	Cold water flow rate
6	CO2 concentration
7	Medium volume
8	PH
9	Fermenter temperature

Table B5 penicillin fermentation process fault types

Fault number	Fault type	Simulation time (h)
1	Base flow rate down	400
2	Agitator power down	400
3	Air flow down	400

Here, the FDA and LLEDA methods are selected for comparison. Figure B7 shows the classification results for each method. Table B6 shows the fault type recognition rates for different methods. Obviously, we can find that the parametric t-SNE has a better classification effect, and it has less uncertainty. From table B6, the parametric t-SNE has a higher accuracy. Therefore, we found that the parametric t-SNE method has better fault classification ability, and the parametric t-SNE can better distinguish non-Gaussian nonlinear industrial fault data with only a small number of features.

Table B6 The correct recognition of faults 1, 2, and 3 by different methods

Number of features	Recognition rate	FDA	LLEDA	parametric t-SNE
d=2	Fault 1	0.965	0.952	1
	Fault 2	0.940	1	1
	Fault 3	0.913	1	1

References

1 Maaten L V D. Learning a parametric embedding by preserving local structure. J Mach Learn Res, 2009, 5: 384-391

2 Maaten L V D, Hinton G. Visualizating data using t-SNE. J Mach Learn Res, 2008, 9: 2579-2605

3 Kouropteva O, Okun O, Pietikäinen M. Selection of the optimal parameter value for the locally linear embedding algorithm. In: the 1^{st} International Conference on Fuzzy Systems and Knowledge Discovery. Singapore, 2002. 359-363

4 Wang R X, Wang J, Zhou J L, et al. Fault diagnosis based on the integration of exponential discriminant analysis and local linear embedding. Can J Chem Eng, 2018, 96: 463-483

5 Wang J, Liu L, Cao L, et al. Fault diagnosis based on kernel Fisher envelope surface for batch processes. CIESC Journal, 2014, 65: 1317-1326

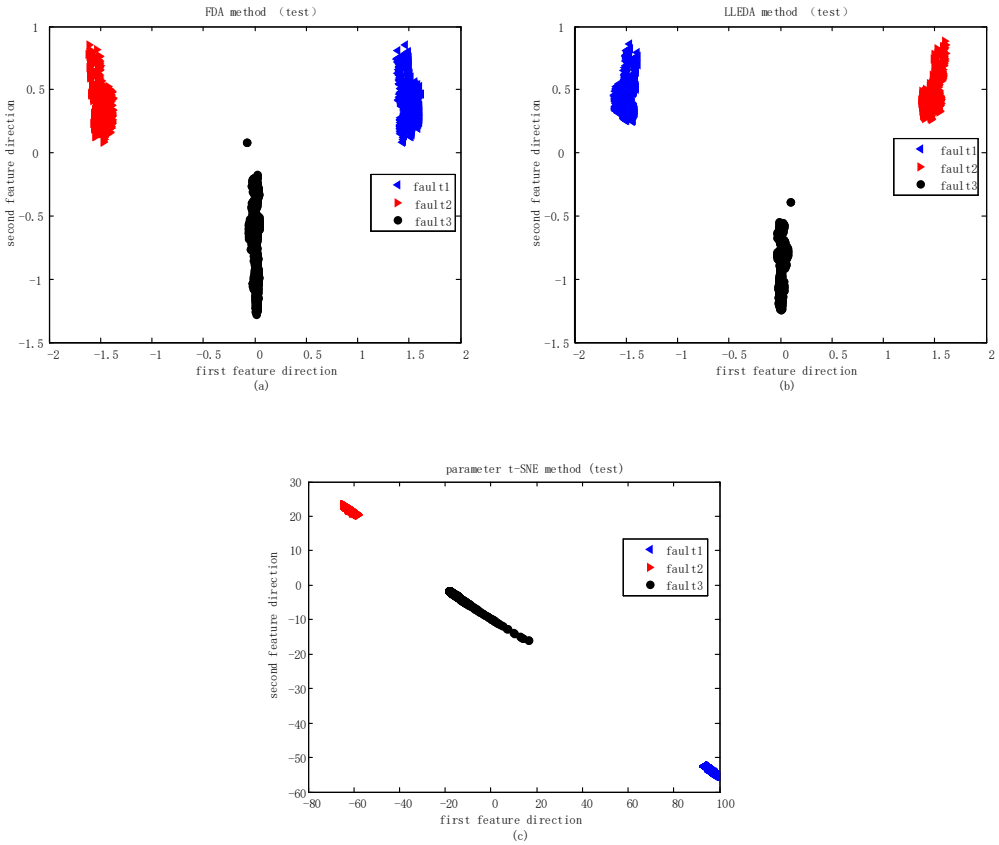

Figure B7 Visualization of 1000 failure data from the fault 1,2,3 using (a) FDA; (b) LLEDA; (c) parametric t-SNE.