processes

MDPI

Article

Fault Diagnosis Using Dynamic Principal Component Analysis and GA Feature Selection Modeling for Industrial Processes

Chenpeng Liu, Jianjun Bai and Feng Wu *

Automation College, Hangzhou Dianzi University, Hangzhou 310018, China
* Correspondence: fengwu@hdu.edu.cn

Abstract: With the continuous expansion of industrial production scale, most of the chemical process variables are nonlinear, multi-modal and dynamic. For some traditional multivariate statistical monitoring and fault diagnosis algorithms, such as principal component analysis (PCA), the premise of its application is that the process data is time-independent. To this end, a dynamic principal component analysis (DPCA) method is proposed. However, since the input matrix of DPCA fault diagnosis needs to add an augmented matrix to the original data matrix, the number of eigenvalues of the augmented matrix is too large and there are many redundant eigenvectors. Therefore, this paper proposes a fault diagnosis and monitoring algorithm combining feature selection and DPCA, which considers the dynamic characteristics of multivariate data and reduces the dimension of the input matrix. At present, the average modeling and diagnostic accuracy of PCA-based fault diagnosis on T^2 statistic is 65.49%, and that on Q statistic is 76.78%. The average modeling and diagnostic accuracy of fault diagnosis based on DPCA on T^2 statistic is 63.17%, and the average modeling and diagnostic accuracy on Q statistic is 83.65%. Finally, through a TE simulation process, this paper proves that the accuracy is greatly improved when using the method proposed in this paper compared with PCA and DPCA.

Keywords: fault diagnosis; PCA; DPCA; genetic algorithms; feature selection

Citation: Liu, C.; Bai, J.; Wu, F. Fault Diagnosis Using Dynamic Principal Component Analysis and GA Feature Selection Modeling for Industrial Processes. *Processes* **2022**, *10*, 2570. https://doi.org/10.3390/pr10122570

Academic Editor: Jean-Pierre Corriou

Received: 27 October 2022
Accepted: 24 November 2022
Published: 2 December 2022

1. Introduction

Due to the rapid development of science and technology, the scale of modern industrial processes is also expanding. During the operation of the chemical process, if the equipment fails or is abnormal and is not repaired in time, extreme conditions, such as high temperature and high pressure, toxic gas release, and even explosion, may occur [1]. This will not only cause huge economic losses to the factory, but also greatly endanger the lives of workers. The emergence of fault diagnosis and monitoring technology has greatly ensured high-quality production in the chemical industry and reduced the frequency of safety accidents. At present, the mainstream research methods of fault diagnosis and monitoring technology mainly include data-based multivariate statistics and artificial intelligence-based deep learning methods [2–5].

Process monitoring technology based on multivariate statistics mainly reduces the dimension of offline data and online data generated in the production process through multivariate statistical analysis. Deep learning mainly uses the powerful learning ability of the computer, the multi-layer hidden layer network, and the layer-by-layer learning method to extract the main information from the input data, and, through a lot of training, to improve the fault accuracy diagnosis [6]. In recent years, many scholars at home and abroad have proposed a variety of fault diagnosis methods and achieved good results. However, industrial process monitoring and fault diagnosis remain a challenging problem due to the large number of complex properties in industrial processes, such as dynamics, nonlinearity, multimodality, and heavy coupling [7–10].

Common multivariate statistics-based methods include principal component analysis (PCA) [11], linear discriminant analysis (LDA) [12], independent component analysis

(ICA) [13], and partial least squares (PLS) [14,15]. Aiming at the dynamic features appearing in chemical process data, Ku et al. proposed a dynamic principal component analysis (DPCA) method by extending the time augmentation matrix [16]. To address the nonlinear features present in the data, Kernel Principal Component Analysis (KPCA) is proposed. This method uses a nonlinear mapping function to map the original data space to a higher-dimensional space, and then performs dimensionality reduction processing through PCA [17,18].

Most of the dimensionality reduction algorithms belong to feature extraction. Through the dimensionality reduction matrix, a new set of feature variables is obtained, which changes the original feature space, often resulting in a loss of some important feature variables in the original feature space. Feature selection, on the other hand, selects a subset of the original feature space that best represents the original feature space without destroying the original data. For data with too large a latitude, finding the optimal feature subset can be regarded as an optimal problem. Through the heuristic optimization algorithm, each feature subset can be uniformly evaluated, and the optimal solution can be finally obtained.

Aiming at certain shortcomings of some existing algorithms, this paper proposes a fault diagnosis method that is built on GA feature selection based on dynamic principal component analysis (GA-DPCA). The main improvements of this paper are as follows:

(1) First, the feature selection based on genetic algorithm is adopted, which can quickly find the optimal feature subset, which not only refines the original data, but also loses the integrity of the data as little as possible.
(2) The data matrix after feature selection is extended with time delay and the residual space of feature selection is added, which effectively considers the autocorrelation and integrity of the data.
(3) Sliding window filtering technology is adopted to remove the noise existing in the data itself.
(4) An experimental comparative study of the proposed method and existing methods is carried out using the Tennessee Eastman process and the actual coking process.

2. PCA and DPCA

PCA is the most widely used data dimensionality reduction algorithm. The main idea of PCA is to map n-dimensional features onto k-dimensions, which is a brand-new orthogonal feature called principal components. It is a k-dimensional feature reconstructed on the basis of the original n-dimensional feature. After mapping, only the features with large variance are selected, and the features with almost zero variance are ignored to achieve dimensionality reduction of the original data features.

Assuming the normalized data are $X = [x_1, x_2, \ldots x_n]^T \in R^{n \times m}$, the formula has n variables and m samples. $\vec{p}$ is the projection vector, $y_i = \vec{p}^T x_i$, and the objective function of PCA is:

$$\begin{aligned} & J_{PCA}\left(\vec{p}\right) = \max\left(\tfrac{1}{n}\sum_i (y_i)(y_i)^T\right) \\ & s.t. \\ & \quad \vec{p}^T \vec{p} = 1 \end{aligned} \tag{1}$$

The data formula can be decomposed into:

$$X = TP^T + E \tag{2}$$

$$C = \frac{1}{n} XX^T \tag{3}$$

where X is the normalized data; $P \in^{m \times k}$ is the loading matrix, which can be obtained by eigen decomposition of C; $T^{n \times k}$ is the score matrix; and E is the residual matrix.

k represents the number of principal components, which can be generally calculated by the cumulative variance contribution rate [19]:

$$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{m} \lambda_i \times 100\% \geq \sigma \tag{4}$$

where λ_i is the eigenvalue of the C eigenvalue decomposition and arranges it from large to small, and σ is the ratio of the sum of the maximum eigenvalues to all the eigenvalues, generally with a value of 85%.

Fault diagnosis mainly uses two statistics, T^2 and Q, to judge whether a fault has occurred. T^2 statistic, also known as Hotelling's T^2, mainly reflects the changes of the spatial characteristics of the principal components, which are defined as follows [20]:

$$T^2 = x_i^T P S^{-1} P^T x_i \tag{5}$$

The Q-statistic, also known as the Squared Prediction Error Index (SPE), mainly reflects the changes in the subspace characteristics of residuals, which are defined as follows:

$$SPE = e^T e \tag{6}$$

The control limits for T^2 and Q statistics are D_C and $Q_{C,a}$, respectively, and the formula is as follows:

$$D_C \sim \frac{l(n^2-1)}{n(n-1)} F_\alpha(l, n-l) \tag{7}$$

$$\begin{aligned} Q_{C,\alpha} &= \theta_1 \left[1 - \theta_2 h_0 \left(\frac{1-h_0}{{\theta_1}^2} \right) + \frac{\sqrt{z_\alpha (2\theta_2 {h_0}^2)}}{\theta_1} \right]^{1/h_0} \\ \theta_i &= \sum_{j=l+1}^{m} \lambda_j^i (i = 1, 2, 3) \\ h_0 &= 1 - \frac{2\theta_1 \theta_3}{3{\theta_2}^2} \end{aligned} \tag{8}$$

Among them, n is the number of modeled data samples; l is the number of principal components retained in the principal components; α is the level of significance; the critical value of F distribution under $l, n-l$ is found in the statistical table; C_α is the critical value of the normal distribution at the significant level α; and λ_j is the eigenvalue of the smaller data covariance matrix.

The DPCA algorithm requires time-delay expansion of the original data space matrix. Assuming that the current data are related to observations from previous L-step data, the expansion matrix is [21]:

$$X(L) = [X(t), X(t-1), \ldots, X(t-L)] \tag{9}$$

3. GA-DPCA

GA feature selection can quickly obtain the optimal feature subset of the original data matrix and reduce the original data without destroying it. The DPCA algorithm can take into account the dynamic characteristics of the data and further reduce the dimension of the data matrix at the same time. Therefore, this paper proposes a dynamic principal component analysis method based on GA feature selection for process monitoring and fault diagnosis through TE datasets. The method flow is shown in Figure 1. The first part is sliding window denoising, the second part is feature selection based on a genetic algorithm, the third part is offline modeling and determination of control limits, and the last is online monitoring.

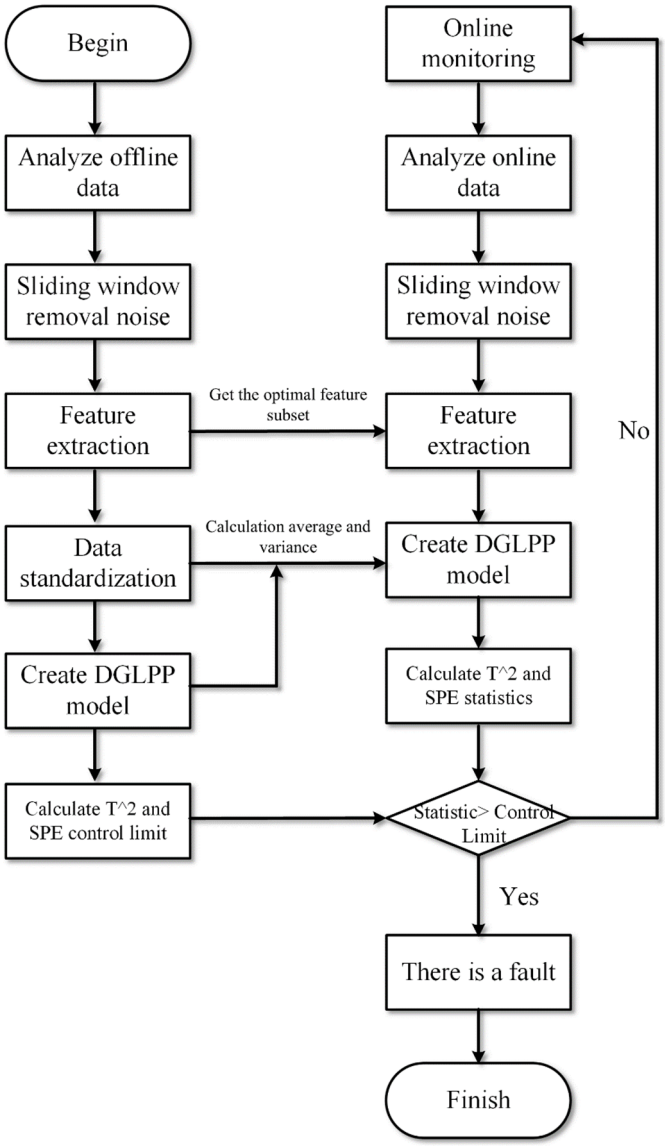


Figure 1. Framework diagram of the whole control system.

3.1. Sliding Window Removal Noise

In industrial process production, data acquisition is often accompanied by a large amount of noise interference, which makes the acquired data inaccurate, thus affecting the accuracy of modeling. In order to suppress noise interference and improve modeling accuracy, we need to use appropriate filtering algorithms to filter the data. This paper mainly adopts the filtering algorithm of sliding window denoising [22]. As shown in Figure 2, the basic idea is to set a sliding window with a fixed width, slide along the time series, and use the arithmetic mean of the data in the window as the output value of the new series, that is, the filtered series. N is set to the width of the sliding window. If $N = 2k + 1$, and the input and output are x_n and y_n, respectively, then:

$$y_n = \frac{1}{2k+1} \sum_{i=-k}^{i=k} x(n+i) \tag{10}$$

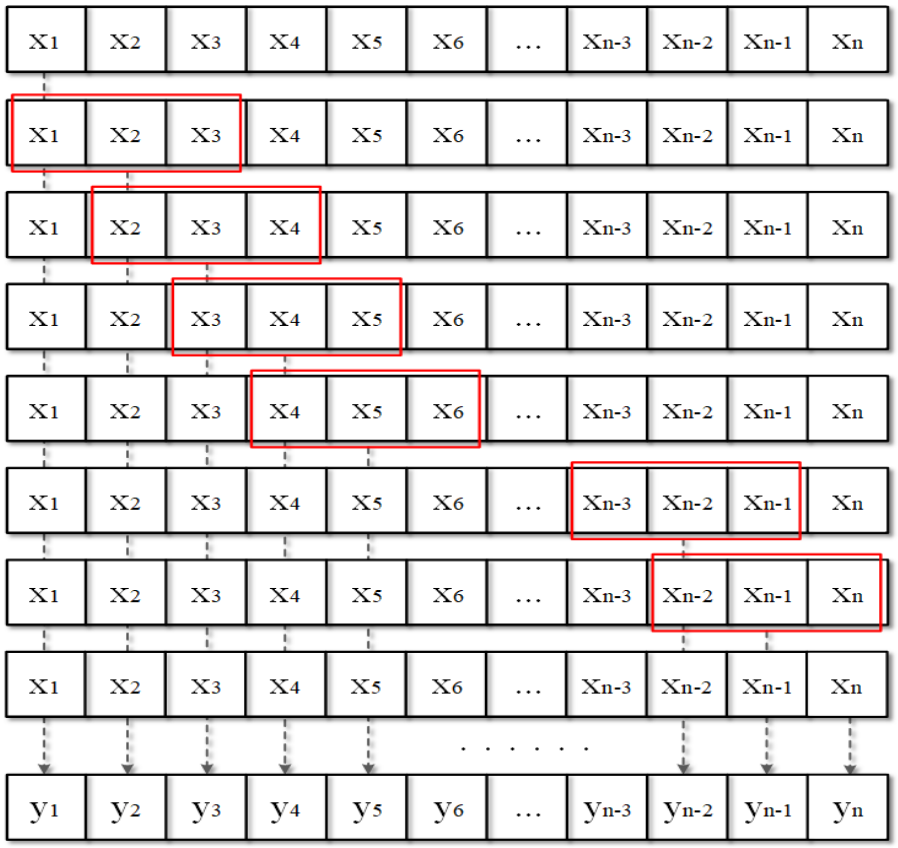


Figure 2. Sliding window denoising diagram.

A piece of data is taken from the training set of the TE dataset and filtered through a sliding window. The results are shown in Figure 3, which shows that the sliding window filtering algorithm can effectively suppress high-frequency noise.

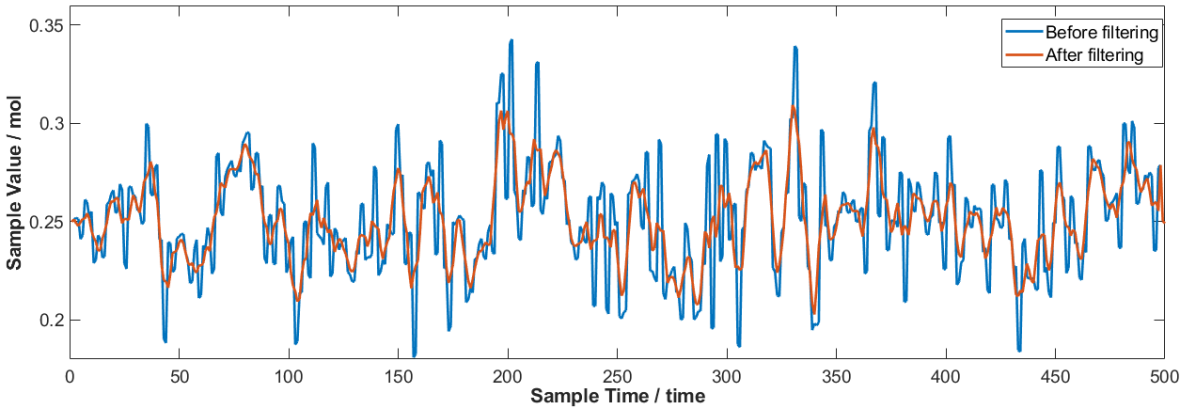


Figure 3. Comparison diagram before and after filtering.

3.2. Feature Selection Based on GA

Feature selection is an important question in feature engineering. Its goal is to find the optimal feature subset. Feature selection can eliminate irrelevant or redundant features, thereby reducing the number of features and improving model accuracy. According to the form of feature selection, it can be divided into three categories: filtering, encapsulation and embedding.

The general process of feature selection is: (1) generating subsets; (2) evaluating functions; (3) stopping criteria; (4) validation process [23,24]. For high-latitude data,

finding the optimal solution can take a lot of time. Genetic algorithm is an adaptive global optimization search algorithm, which is a computational model that simulates the natural selection and genetic mechanism of Darwin's theory of biological evolution. Therefore, the feature selection based on genetic algorithm will greatly increase the speed of obtaining the optimal solution. As shown in Figure 4, the general steps of GA-based feature selection are mainly divided into the following steps: initializing the population, calculating the fitness, setting the genetic operator, and updating the iterations [25,26].

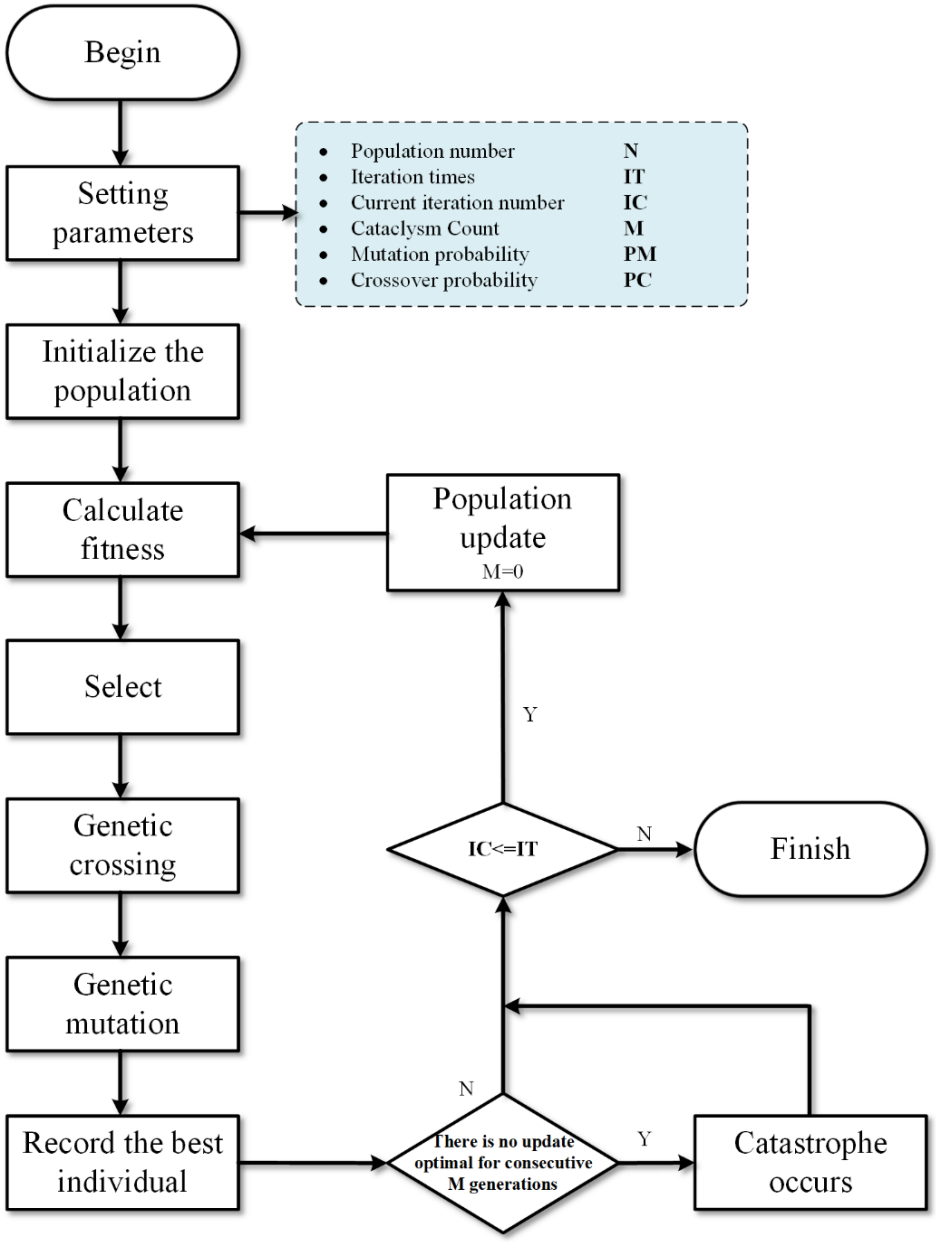


Figure 4. Genetic algorithm flow chart.

3.2.1. Initializing the Population

Genetic algorithm follows the law of survival of the fittest and elimination of the weak, that is, retaining individuals with strong fitness and eliminating individuals with weak fitness. Before starting the genetic algorithm, there are several parameters that need to be set in advance: the first is the size of the population. If the population size is too small, inbreeding will obviously occur, resulting in genetic disease. Conversely, if the population size is too large, the results will be difficult to converge, resources will be wasted, and

the robustness will be reduced. The second parameter is the number of iterations. If the evolutionary algebra is too small, the algorithm is not easy to converge, and the population is immature. If the evolutionary algebra is too large, the population may converge too early, and it is pointless to continue evolution, which would only increase time consumption and waste resources. Further, there is the probability of mutation. If the probability of mutation is too small, the diversity of population genes cannot be guaranteed. If the probability of mutation is too large, the genetic algorithm will degenerate into random search. Similar to the mutation probability, if the mating probability is too large, the existing beneficial individuals will be destroyed, the randomness will increase, and the optimal individual will be easily lost. If the mating probability is too small, the population cannot be effectively updated. The last is the catastrophe count. When the population has not updated the optimal solution for N consecutive generations, it is necessary to set the catastrophe to prevent the population from falling into the local optimal solution. If the value of N is too small, the search path will be greatly increased or even impossible. If the value of N is too large, the function of the mutation operator will become small or even useless, so an appropriate value should be set.

Genetic coding is required before an individual can be produced. The coding method directly affects the operation of genetic operators, such as the crossover operator and mutation operator of the genetic algorithm, thus determining the efficiency of genetic evolution to a large extent. In this post, feature retention and loss is essentially a 0–1 problem. Assuming that the processed data are $X = [x_1, x_2, \ldots, x_n]$, where x_n is the feature of the data, if 1 means that this eigenvector is retained, then 0 means rounding off that eigenvector, the entire genotype of the individual is a binary encoded symbol string. For binary encoding, its advantages are that it is simple, it easily implements operations such as crossover and mutation, and it conforms to the principle of minimum character set encoding. The algorithm can be theoretically analyzed by using the mode theorem. Since the feature selection problem belongs to the type of planning problem, it is very suitable for binary encoding.

3.2.2. Calculating Fitness

Individual fitness refers to the degree of dominance of an individual in the survival of the population, and is used to distinguish the "good" and "bad" individuals. The fitness of an individual can be calculated using the fitness function. The fitness function, also known as the evaluation function, is mainly used to judge the fitness of an individual through individual characteristics. The general process for evaluating individual fitness is as follows:

1. After decoding the individual code string, the individual phenotype can be obtained.
2. The objective function value of the corresponding individual can be calculated from the individual's phenotype.
3. According to the type of optimization problem, the fitness of individuals can be calculated from the objective function value according to certain transformation rules.

The purpose of feature selection is to extract the most representative feature subset from the original feature vector data. This subset should satisfy the following two conditions as much as possible:

The correlation between the feature subset vectors after feature selection should be as small as possible.

The correlation between the feature subset after feature selection and the residual data matrix should be as large as possible.

Assuming that the selected data of an individual feature are $X_1 = [x_1, x_2, \ldots, x_m]$, the data in the residual space of the individual is then $X_2 = [x_1, x_2, \ldots, x_k]$, $k + m = n$. If there

are two groups of data, $A = \{a_1, a_2, \ldots, a_N\}$, $B = \{b_1, b_2, \ldots, b_N\}$, the Pearson correlation coefficient shows that the correlation coefficients of these two groups of data are:

$$\rho_{AB} = \frac{Cov(A,B)}{\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n} \frac{(A_i - E(A))}{\sigma_A} \frac{(B_i - E(B))}{\sigma_B}}{n}$$
$$\sigma_A = \sqrt{\frac{\sum_{i=1}^{n} (A_i - E(A))}{n}}, \sigma_B = \sqrt{\frac{\sum_{i=1}^{n} (B_i - E(B))}{n}} \tag{11}$$

where σ_A, σ_B is the standard deviation of A, B, respectively. At the same time, it is obvious that $|\rho_{AB}| \leq 1$. The fitness function is as follows:

$$f_{Fit} = p \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} \left| \rho_{X_1(i) X_1(j)} \right|}{\sum_{i}^{m-1} i} + (1-p) \sum_{i=1}^{m} \sum_{j=1}^{k} \left| \rho_{X_1(i) X_2(j)} \right| \tag{12}$$
$$i = 1, 2, \ldots, m$$
$$j = 1, 2, \ldots, k$$

where f_{Fit} is the fitness function, and p is the fitness proportion of the above two indicators, $0 \leq p \leq 1$.

3.2.3. Setting up Genetic Strategies

The genetic operators of the genetic algorithm proposed in this paper include the elite selection operator, gene crossover operator, gene mutation operator and mutation operator. The elite retention strategy [27,28] aims to prevent the loss of the best individuals in the current population in the next generation, thereby preventing the genetic algorithm from converging to the global optimal solution. The specific process is to directly copy the best individual currently existing in the population to the next generation and replace the worst individual in the next generation when the individual is selected. The elite retention strategy improves the global convergence of the standard genetic algorithm, and it is theoretically proved that the standard genetic algorithm with elite retention has global convergence. Catastrophic strategies [29,30] aim to further improve global search performance. Its essence is to simulate the biological evolution process. When the external environment undergoes major changes, such as volcanoes and tsunamis, biological populations will face mass death or even extinction.

(1) Elite selection strategy

The most common selection operator in genetic algorithms is the roulette method, in which the probability of selecting an individual is directly proportional to its fitness value. Essentially, the strategy of elite retention is essentially the same as that of keeping individuals as well-adapted as possible. The idea is to have the most adapted individuals directly retain and replicate their offspring while eliminating the less adapted ones, which converges faster than roulette. In the genetic algorithm proposed in this paper, the specific strategies for elite preservation are to retain the first 25% of the fitness of the individuals and duplicate them to enter the next generation directly, retain 25%–50% of the individuals to enter the next generation, and directly eliminate the last 50%.

(2) Gene crossover strategy

Crossover is the main process of generating new individuals in genetic algorithms. It exchanges some chromosomes between two individuals with a certain probability. This paper adopts the method of single-point chromosome segment crossover. The specific procedure is as follows: first, the populations are paired randomly; second, the position of the intersection is randomly set. Finally, some genes are exchanged between pairs of chromosomes. As shown in Figure 5, at the n-1 position of the chromosome, individual A crosses with individual B, where A changes from 0 to 1 and B changes from 1 to 0.

(3) Gene mutation strategy

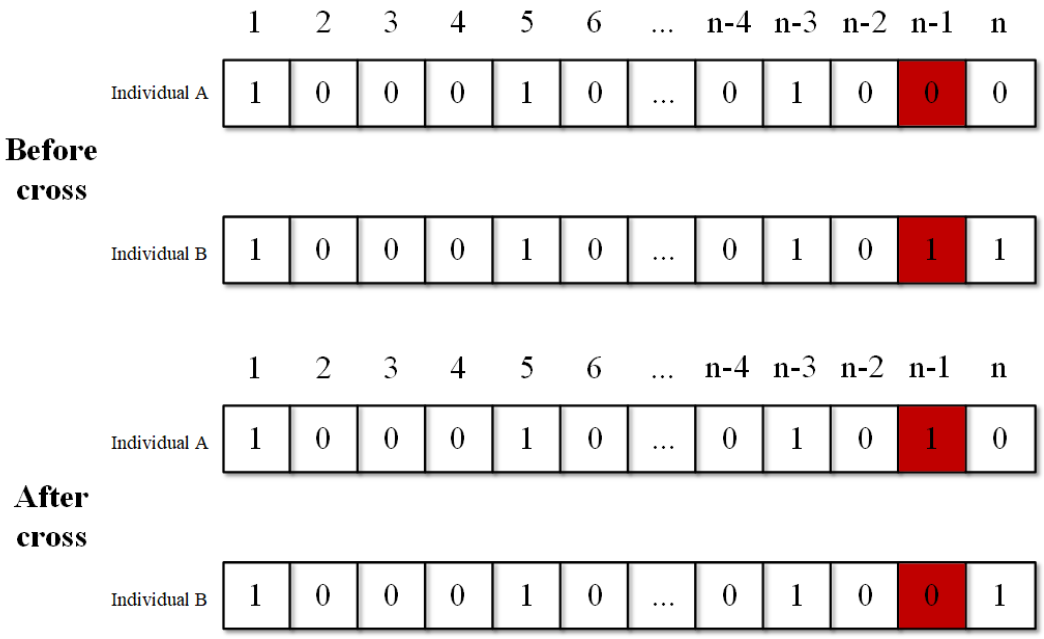


Figure 5. Gene crossover diagram.

The mutation operation is to change the value of a gene with a small probability at one or a certain locus of an individual. It reflects the diversity of genes. In this paper, we use the single-point mutation method to perform mutation operations. The specific operation process is to first determine the mutation position of each individual. Then the original gene value of the mutation point is inverted according to a certain probability. As shown in Figure 6, the gene of individual A at the n-1 position mutates from 0 to 1.

(4) Catastrophe Strategy

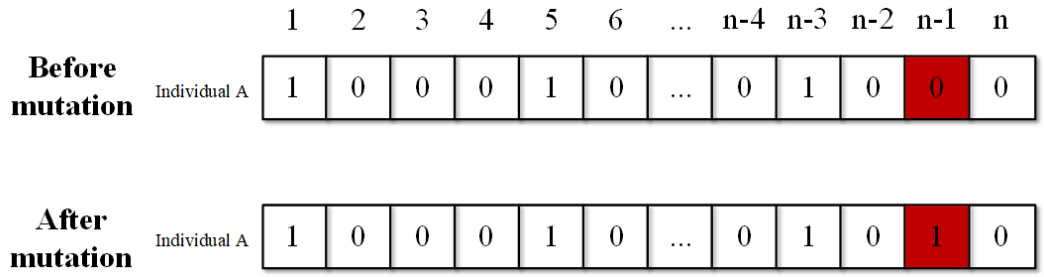


Figure 6. Gene mutation diagram.

The genetic algorithm has strong local search ability, but it is easy to fall into local optimal solution. In order to jump out of the local optimum, all excellent individuals must be killed, so that there is enough evolution space for points far from the current extreme value. This is the idea of cataclysm. Holocaust is killing the best individuals, thereby producing more good species. A countdown to when the disaster will happen is required. In the genetic algorithm of this paper, if there is not a better individual than before in n consecutive generations, a disaster occurs, and the disaster makes the count n equal to 0.

3.2.4. Update Iteration

After one selection and cross-mutation, if the current disaster count does not reach the threshold, it will directly enter the next generation, otherwise, the disaster operation will

be performed, and the disaster count will return to 0 and enter the next generation. Finally, if the current algebra reaches the number of iterations, the iteration is stopped.

3.3. Offline Modeling

The steps of offline modeling and monitoring are as follows:

(1) Collect the data of normal operation of the chemical process and conduct sliding window denoising on the data to obtain X_{train}. Then, due to the inconsistency of different variable units of process data, the data needs to be standardized. The standardization process is as follows:

$$\widetilde{X}_{train} = \frac{X_{train} - \overline{X}_{train}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_{train} - \overline{X}_{train}\right)^2}} \tag{13}$$

where $\overline{X}_{train}$ represents the mean matrix established by the mean of each variable in the data and $\widetilde{X}_{train}$ represents the normalized process data matrix.

(2) Overlay the standardized data with observations from previous L moments to construct a new data matrix, $X_{Train} = \left[\widetilde{X}_{train,1}(0), \widetilde{X}_{train,1}(1), \ldots, \widetilde{X}_{train,1}(L), \widetilde{X}_{train,2}\right]$, for DGLPP. L is the delay parameter, which is generally 1 or 2 depending on the actual situation. $\widetilde{X}_{train,1}(L)$ represents data delayed to L steps and selected for features. $\widetilde{X}_{train,2}$ represents residual data for feature selection.

(3) Establish a monitoring model based on DPCA according to Formulas (1)–(4).

(4) Calculate T^2 and Q statistics according to Formulas (5) and (6).

3.4. Determining Control Limits

The control limits are established to determine whether a failure has occurred at any point in time, so the control limits for the T^2 and Q statistics need to be calculated. When the statistic value calculated by the online collected data is greater than the control limit, it is regarded as a failure. The control limit formulas for determining the statistics are shown in Formulas (7) and (8).

3.5. Online Monitoring

The steps of online monitoring are as follows:

(1) Collect abnormal data in the process of chemical industry and obtain X_{test} by sliding window denoising, then standardize the data by using Formula (12). Finally, the new data sample required for DGLPP is constructed as $X_{Test} = \left[\widetilde{X}_{test,1}(0), \widetilde{X}_{test,1}(1), \ldots, \widetilde{X}_{test,1}(L), \widetilde{X}_{test,2}\right]$.

(2) Using the projection matrix A obtained from the offline process and the newly acquired data, an online monitoring model based on DPCA is established:

$$\begin{aligned} X_{Test} &= Ay_{Test} + e_{Test} \\ y_{Test} &= (A^T A)^{-1} A^T X_{Test} \\ e_{Test} &= X_{Test} - Ay_{Test} \end{aligned} \tag{14}$$

where y_{Test} is the vector after projection, e_{Test} is the residual vector, and A is the projection matrix.

(3) Calculate T^2 and Q statistics according to Formulas (5) and (6).

(4) Compare with the control limit established in offline modeling to determine whether there is a fault.

3.6. Fault Diagnosis

After statistics are found to be abnormal, fault diagnosis is required. In this paper, the improved contribution graph is used for fault diagnosis, and the formula is as follows:

$$cont_j = \sum_{i=1}^{d} \frac{{t_{j,i}}^T t_{j,i}}{\lambda_\eta(j,i)} \tag{15}$$

$$q_j = {e_j}^2 \tag{16}$$

where d is the number of principal elements, j is the number of variables, and e_j represents the residual value of element j.

Considering the dynamic nature of the data, traditional contribution graphs cannot be used directly. The contribution diagram of the improvement used in this paper is as follows:

$$\begin{gathered} CONT_j = cont_j + cont_{j+m} + cont_{j+2\times m} + \ldots + cont_{j+L\times m}\ j \le m \\ Q_j = q_j + q_{j+m} + q_{j+2\times m} + \ldots + q_{j+L\times m}\ j \le m \end{gathered} \tag{17}$$

Here, m is the number of variables in the original data before matrix expansion, and the variable with the largest result is the cause of the failure.

4. Simulation

In order to prove the effectiveness of the algorithm proposed in this paper, we used the Eastman process simulation data in Tennessee to verify it, and compared the PCA and DPCA algorithms to show the superiority of the algorithm. We also set the latency parameter L to 1 with a confidence level of 99%.

TE dataset is produced by Tennessee Eastman, an open and challenging simulation platform for chemical models developed by Eastman Chemical Company in the United States. It has time-varying, strong coupling, and nonlinear properties, and has been widely used to compare various monitoring methods [31]. The four reactants, A, C, D, and E, and inert B are fed into the reactor where products G and H are formed, and by-product F is formed [32]. The control structure is shown in Figure 7. The process has 22 continuous process measurements, 19 component measurements, and 11 control variables for a total of 52 variables. Table 1 lists the 52 variables and Table 2 describes the failure of the 21 sets of test data. The training and testing datasets for each failure are represented by 480 and 960 observations, respectively.

The proposed GA-DPCA is compared with PCA, DPCA. The cumulative variance contribution method is used to determine the number of principal components. The control limits for all monitoring statistics for the three methods were set to 99%. The monitoring results of the three algorithms PCA, DPCA and GA-DPCA are shown in Table 3. It can be seen from Table 3 that the average diagnostic rate of the T^2 statistic of GA-DPCA is higher than that of the other two algorithms, and the Q statistic of GA-DPCA is slightly lower than that of DPCA, but higher than that of PCA. In the cases of faults 4, 11, 17, 18 and 21, the diagnostic rate of the T^2 statistic of GA-DPCA was significantly higher than that of the other two methods. In the cases of faults 5, 10, 11 and 20, the diagnostic rate of the Q statistic of GA-DPCA is significantly higher than that of the other two methods. In other cases, the diagnostic rate of GA-DPCA is about the same as the other two algorithms. These situations reflect that the GA-DPCA algorithm has a significant improvement in the diagnosis rate of these small faults that are not easily detected by PCA or DPCA. In order to further illustrate the superiority of the GA-DPCA monitoring method, the comparison of the detection results of PCA, DPCA and GA-DPCA algorithms in faults 4 and 11 is shown in Figures 8 and 9, respectively.

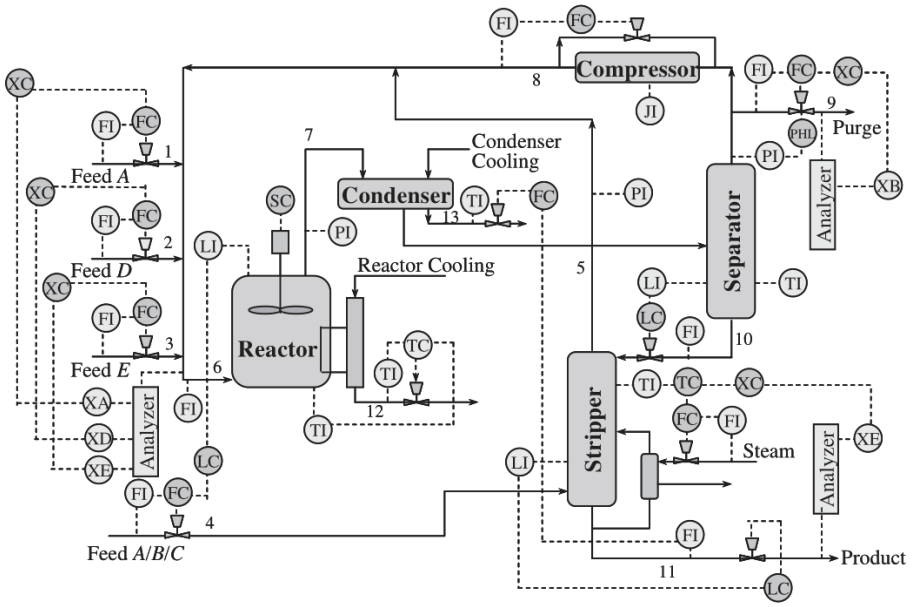


Figure 7. TE process.

Table 1. Process monitoring variables in the Tennessee Eastman process.

No.	Variables	No.	Variables
1	A feed (stream 1)	27	Ingredient E (stream 6)
2	D feed (stream 2)	28	Ingredient F (stream 6)
3	E feed (stream 3)	29	Ingredient A (stream 9)
4	Total feed (stream 4)	30	Ingredient B (stream 9)
5	Recycle flow (stream 8)	31	Ingredient C (stream 9)
6	Reactor feed rate (stream 6)	32	Ingredient D (stream 9)
7	Reactor pressure	33	Ingredient E (stream 9)
8	Reactor level	34	Ingredient F (stream 9)
9	Reactor temperature	35	Ingredient G (stream 9)
10	Purge rate (stream 9)	36	Ingredient H (stream 9)
11	Product separator temperature	37	Ingredient D (stream 11)
12	Product separator level	38	Ingredient E (stream 11)
13	Product separator pressure	39	Ingredient F (stream 11)
14	Product separator underflow (stream 10)	40	Ingredient G (stream 11)
15	Stripper level	41	Ingredient H (stream 11)
16	Stripper pressure	42	D feed flow valve (stream 2)
17	Stripper underflow (stream 11)	43	E feed flow valve (stream 3)
18	Stripper temperature	44	A feed flow valve (stream 1)
19	Stripper steam Flow	45	Total feed flow valve (stream4)
20	Compressor work	46	Compressor recycle valve
21	Reactor cooling water outlet temperature	47	Purge valve (stream 9)
22	Separator cooling water outlet temperature	48	Separator pot liquid flow valve (stream 10)
23	Ingredient A (stream 6)	49	Stripper liquid product flow valve (stream 11)
24	Ingredient B (stream 6)	50	Stripper steam valve
25	Ingredient C (stream 6)	51	Reactor cooling water flow
26	Ingredient D (stream 6)	52	Condenser cooling water flow

Table 2. Process faults for the TE process.

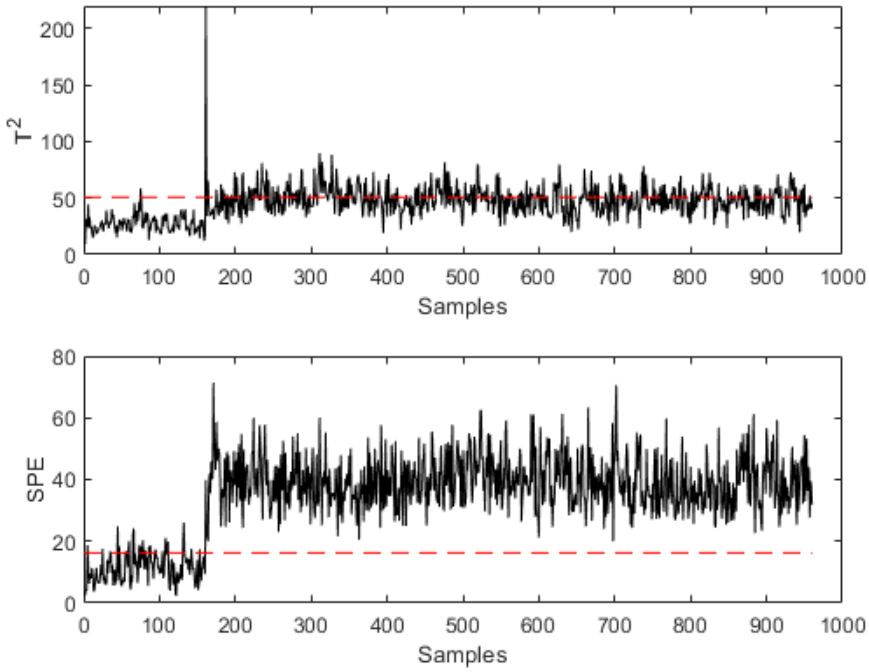
Fault Number	Process Variable	Type
1	A/C feed ratio, B composition constant (stream 4)	Step
2	B composition, A/C ratio constant (stream 4)	Step
3	D feed temperature (stream 2)	Step
4	Reactor cooling water inlet temperature	Step
5	Condenser cooling water inlet temperature	Step
6	A feed loss (stream 1)	Step
7	C header pressure loss-reduced availability (stream 4)	Step
8	A, B, C feed composition (stream 4)	Random variation
9	D feed temperature (stream 2)	Random variation
10	C feed temperature (stream 4)	Random variation
11	Reactor cooling water inlet temperature	Random variation
12	Condenser cooling water inlet temperature	Random variation
13	Reaction kinetics	Slow drift
14	Reactor cooling water valve	Sticking
15	Condenser cooling water valve	Sticking
16	Unknown	Unknown
17	Unknown	Unknown
18	Unknown	Unknown
19	Unknown	Unknown
20	Unknown	Unknown
21	Valve position constant (stream 4)	Constant position

Table 3. Fault diagnosis rate of four methods for the TE Process.

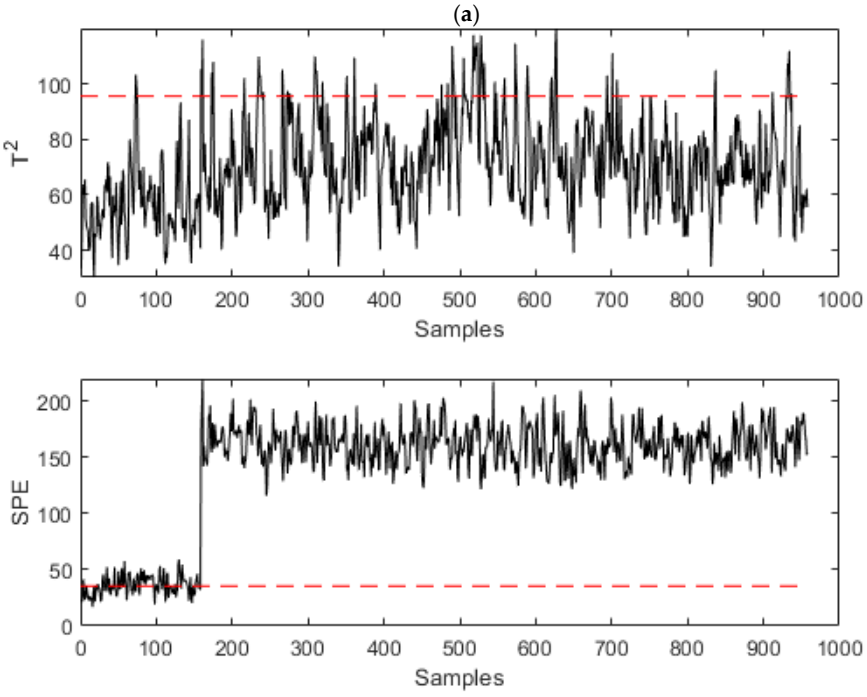
Fault Number	PCA		DPCA		GA-DPCA	
	T^2	Q	T^2	Q	T^2	Q
1	99.58%	98.02%	99.58%	91.14%	99.27%	92.91%
2	98.44%	96.88%	98.75%	91.46%	98.85%	94.79%
3	18.54%	28.75%	17.19%	55.92%	21.79%	46.56%
4	50.73%	97.19%	23.85%	91.77%	93.53%	93.12%
5	38.44%	49.80%	37.29%	77.29%	51.30%	91.97%
6	99.17%	98.54%	99.38%	93.57%	99.79%	95.20%
7	100.00%	98.13%	100.00%	94.20%	99.79%	94.16%
8	97.81%	96.88%	97.71%	92.10%	97.39%	93.01%
9	18.85%	28.33%	16.87%	54.89%	21.48%	50.42%
10	51.36%	74.06%	52.71%	83.57%	64.03%	84.37%
11	60.00%	76.05%	42.39%	91.37%	83.94%	92.60%
12	98.85%	95.84%	99.27%	93.46%	99.06%	94.06%
13	95.94%	95.21%	95.83%	92.32%	96.14%	92.70%
14	99.80%	96.67%	99.90%	90.82%	34.20%	90.83%
15	20.73%	32.08%	19.48%	52.71%	30.03%	46.67%
16	34.90%	68.54%	32.29%	82.51%	45.15%	81.14%
17	83.23%	94.90%	81.67%	90.74%	91.76%	90.20%
18	90.94%	91.98%	91.04%	87.49%	92.18%	88.85%
19	22.29%	56.35%	21.25%	89.60%	26.28%	67.40%
20	47.81%	73.02%	48.75%	84.60%	56.73%	87.29%
21	47.92%	65.21%	51.46%	75.20%	67.47%	71.97%
Average	65.49%	76.78%	63.17%	83.65%	70.01%	82.87%

It can be seen from Figure 8 that in the case of fault 4, for PCA under T^2 statistic monitoring, the effect is good before the fault is introduced, and after the fault is introduced, the T^2 control limit cannot effectively identify the fault. For PCA under the monitoring of the Q statistic, the diagnosis effect before the introduction of the fault is average. After the introduction of the fault, the Q control limit can effectively identify the fault, but it can be

seen from the figure that there is a large fluctuation. For DPCA, the monitoring effect of the T^2 statistic is very poor, and most of the statistical values are below the control limit, almost losing the monitoring effect. For DPCA under the monitoring of the Q statistic, there are errors in the diagnosis effect before and after the fault is introduced. For GA-DPCA under the monitoring of the T^2 statistic, the diagnosis effect is good before and after the introduction of the fault, and the accuracy rate is close to 100%. For GA-DPCA under the monitoring of the Q statistic, there are some errors in the diagnosis before the fault is introduced, but the diagnosis rate after the fault is introduced is 100%.

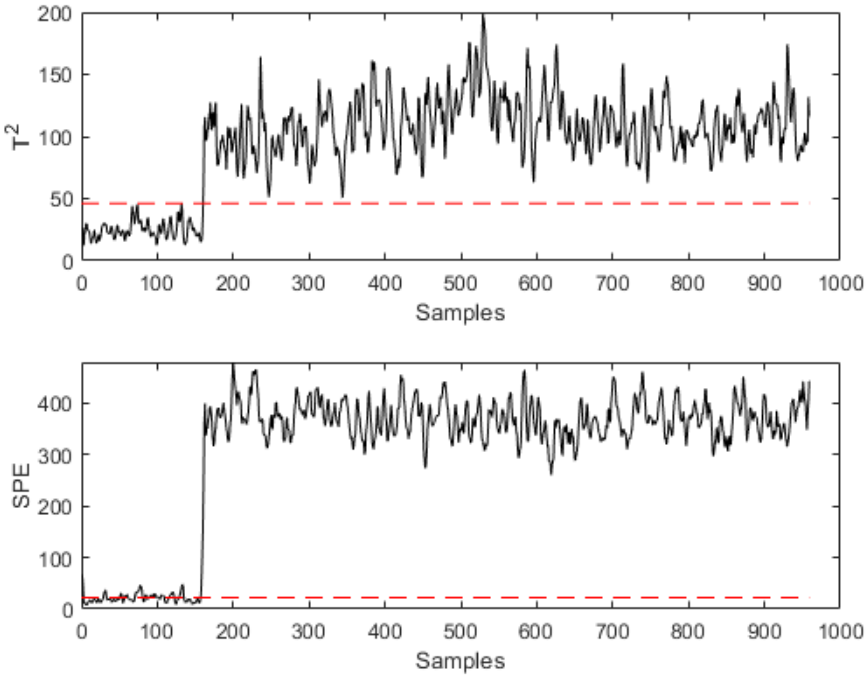


(**a**)



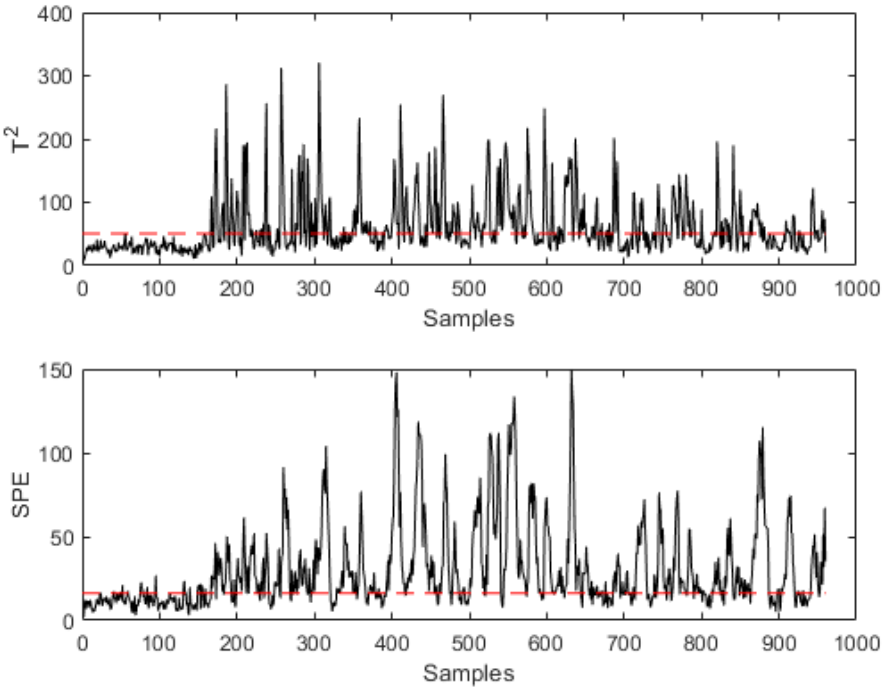
(**b**)

Figure 8. *Cont.*



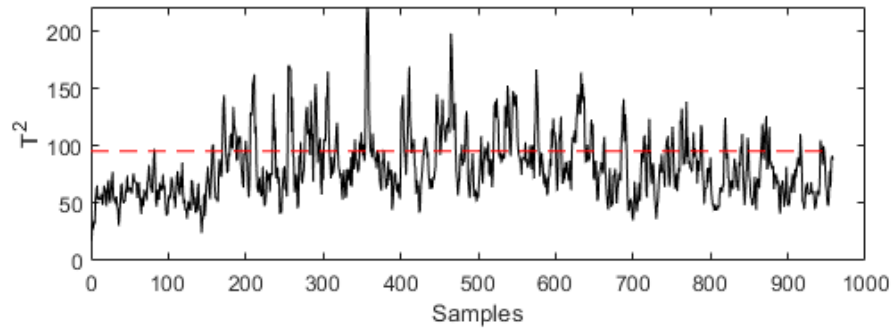
(**c**)

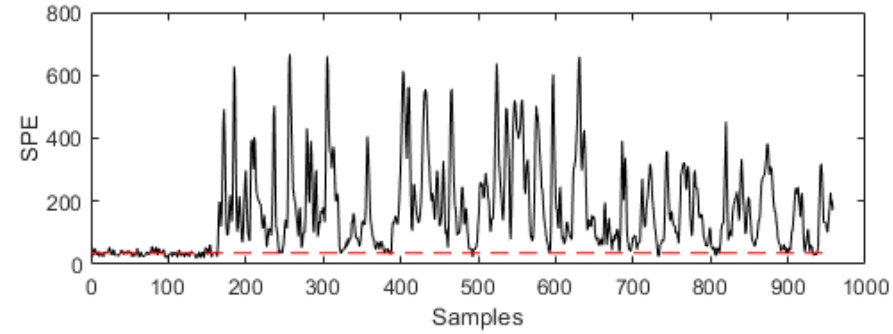
Figure 8. (**a**) Monitoring results of PCA in fault 4. (**b**) Monitoring results of DPCA in fault 4. (**c**) Monitoring results of GA-DPCA in fault 4.



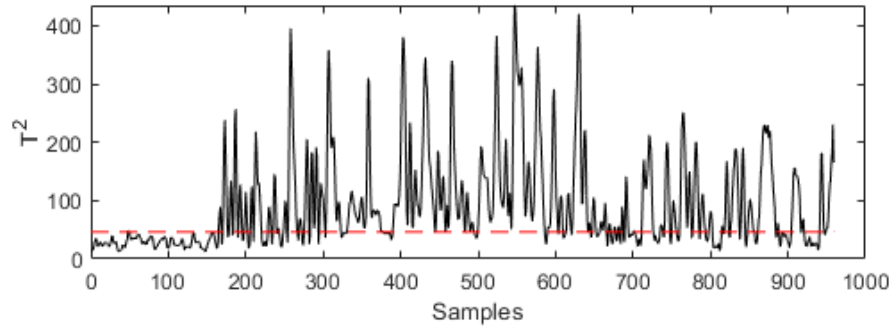
(**a**)

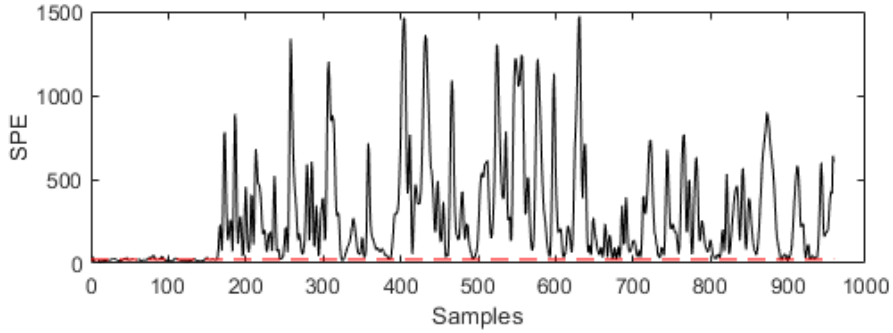
Figure 9. *Cont.*





(**b**)





(**c**)

Figure 9. (**a**) Monitoring results of PCA in fault 11. (**b**) Monitoring results of DPCA in fault 11. (**c**) Monitoring results of GA-DPCA in fault 11.

It can be seen from Figure 9 that in the case of fault 11, for PCA under T^2 statistic monitoring, the effect is good before the fault is introduced, and after the fault is introduced, the T^2 statistic monitoring effect has some errors. For PCA under monitoring of the Q statistic, the diagnosis effect before and after the fault is introduced is average, and the statistics fluctuate greatly. For DPCA under the monitoring of T^2 statistic, the diagnosis effect before the introduction of the fault is good, and after the introduction of the fault, most of the fault data cannot be identification since the diagnosis rate is low. For DPCA

under monitoring of the Q statistic, before the fault is introduced, the misjudgment rate of the diagnostic effect is relatively large, and after the fault is introduced, the accuracy rate is close to 100%. For GA-DPCA under monitoring of the T^2 statistic, the diagnostic effect before and after the introduction of the fault is good, and the accuracy is close to 95%. For GA-DPCA under monitoring of the Q statistic, there are some errors in the diagnosis before the fault is introduced, but the diagnosis rate after the fault is introduced is 100%.

After the fault detection is completed, fault diagnosis was carried out by the contribution diagram. Failure 4 and Failure 11 were selected for testing. The contribution of the Q-statistic based on GA-DPCA to the sample is shown in Figures 10 and 11. In the contribution diagram, the red represents the variable causing the fault, while the other variables are represented in blue. Therefore, variable 51 (reactor cooling water flow) was determined to be the cause of failures 4 and 11. Table 2 was checked to make sure they are correct. Table 4 shows how much GA-DPCA improves the average diagnostic accuracy of the T^2 and Q statistics. Through the TE simulation process simulation results shown in Table 4, the average diagnostic accuracy of GA-DPCA-based fault diagnosis on T^2 statistics is 70.01%, and the average diagnostic accuracy on Q statistics is 82.87%. Compared with PCA, it increased by 4.52% in T^2 and by 6.09% in Q statistics. Compared with DPCA, it increased by 6.84% in T^2 statistics and decreased by 0.78% in Q statistics. Therefore, compared with PCA and DPCA, GA-DPCA has a better monitoring and diagnosis effect in practical applications. In addition, the improved contribution map can accurately diagnose which variables caused the failure, which has practical application value for actual chemical process monitoring.

Table 4. GA-DPCA average improvement rate of accurate diagnosis.

Method	T^2	Q
PCA	4.52%	6.09%
DPCA	6.84%	−0.78%

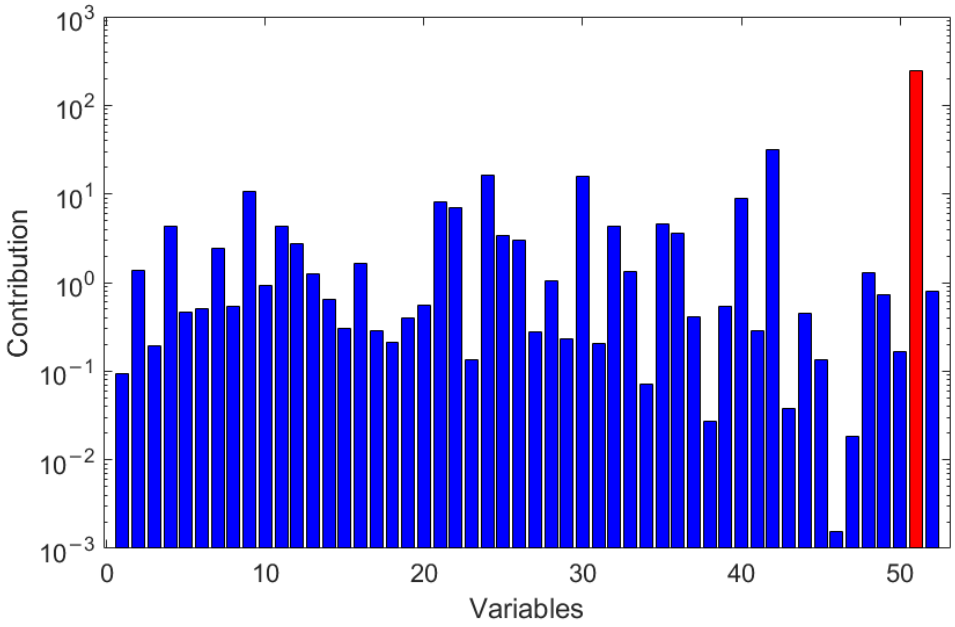


Figure 10. Contribution plots of fault 4.

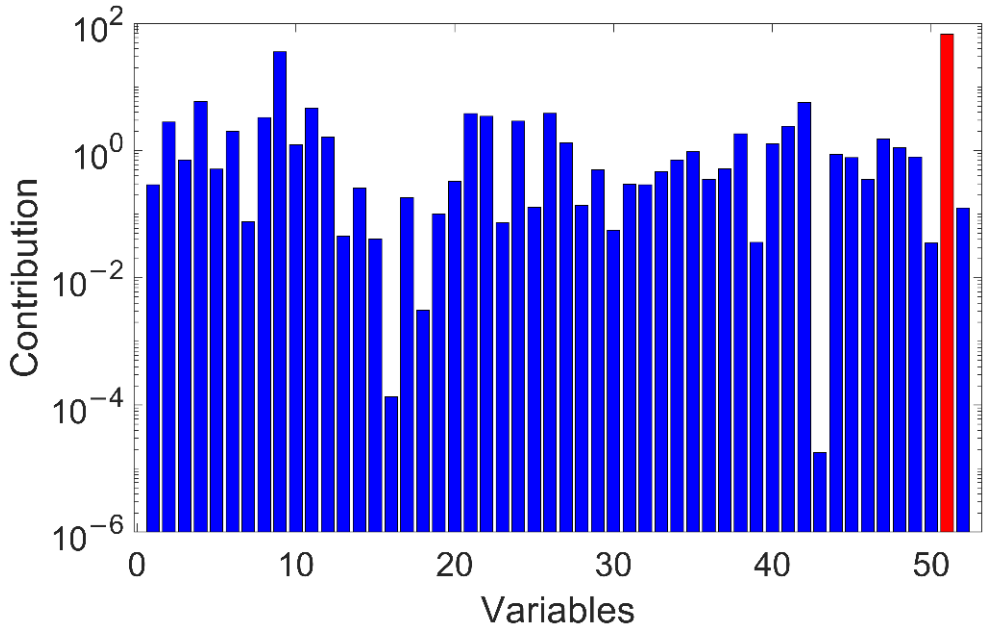


Figure 11. Contribution plots of fault 11.

5. Conclusions

The method proposed in this paper is essentially based on the PCA method. Due to the wide application of the PCA method, many scholars at home and abroad have studied it to improve the monitoring effect. This paper considers the data feature selection aspect and puts forward a DPCA fault monitoring and diagnosis method based on feature selection by using the feature of fast optimal solution of the genetic algorithm. Firstly, sliding window denoising is introduced for removing noise and increasing modeling accuracy. Secondly, feature selection is performed based on a genetic algorithm to find the optimal feature subset and eliminate irrelevant or redundant features. Thirdly, offline modeling and determination of control limits are introduced for process monitoring. This method not only preserves the dimensionality reduction of principal component analysis, but also considers the dynamics and integrity of the data, which can effectively improve the monitoring effect compared with typical methods shown in [33–35].

Although the method presented in this paper provides improved results compared with traditional PCA methods, there is still much room for improvement, for example, diagnostic accuracy under T^2 statistics is still less than 80%. At the same time, due to the non-linear nature of the chemical process data, the effect of traditional PCA is greatly reduced, so the effect of some unknown fault diagnosis is not ideal. In future research, the authors will consider existing algorithms to propose more efficient monitoring methods for chemical processes in view of the non-linear characteristics of high-dimensional data.

Author Contributions: Conceptualization, C.L. and F.W.; methodology, C.L.; software, C.L.; validation, C.L., J.B. and F.W.; formal analysis, F.W.; writing—original draft preparation, C.L.; writing—review and editing, J.B.; visualization, C.L.; supervision, F.W.; project administration, F.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Zhejiang Provincial Natural Science Foundation of China under Grant (LZ22F030001).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, J.; Zhou, D.-H.; Si, X.; Chen, M.-Y.; Xu, C.-H. Review of incipient fault diagnosis methods. *Kongzhi Lilun Yu Yingyong/Control Theory Appl.* **2012**, *29*, 1517–1529.
2. Zhang, R.; Peng, Z.; Wu, L.; Yao, B.; Guan, Y. Fault Diagnosis from Raw Sensor Data Using Deep Neural Networks Considering Temporal Coherence. *Sensors* **2017**, *17*, 549. [CrossRef] [PubMed]
3. Yang, F.; Cui, Y.; Wu, F.; Zhang, R. Fault Monitoring of Chemical Process Based on Sliding Window Wavelet DenoisingGLPP. *Processes* **2021**, *9*, 86. [CrossRef]
4. Liu, K.; Lu, N.; Wu, F.; Zhang, R.; Gao, F. Model Fusion and Multiscale Feature Learning for Fault Diagnosis of Industrial Processes. *IEEE Trans. Cybern.* **2022**, 1–14. [CrossRef] [PubMed]
5. Wang, N.; Yang, F.; Zhang, R.; Gao, F. Intelligent Fault Diagnosis for Chemical Processes Using Deep Learning Multimodel Fusion. *IEEE Trans. Cybern.* **2022**, *52*, 7121–7135. [CrossRef]
6. Lei, Y.; Jiang, W.; Jiang, A.; Zhu, Y.; Niu, H.; Zhang, S. Fault Diagnosis Method for Hydraulic Directional Valves Integrating PCA and XGBoost. *Processes* **2019**, *7*, 589. [CrossRef]
7. Kim, J.; Lee, T.; Lee, S.; Lee, J.; Lee, W.; Kim, Y.; Park, J. A Study on Deep Learning-Based Fault Diagnosis and Classification for Marine Engine System Auxiliary Equipment. *Processes* **2022**, *10*, 1345. [CrossRef]
8. Xu, X.; Feng, J.; Wang, H.; Zhang, N.; Wang, X. Dynamics Analysis of Misalignment and Stator Short-Circuit Coupling Fault in Electric Vehicle Range Extender. *Processes* **2020**, *8*, 1037. [CrossRef]
9. Sun, J.; Xiao, Z.; Xie, Y. Automatic multi-fault recognition in TFDS based on convolutional neural network. *Neurocomputing* **2017**, *222*, 127–136. [CrossRef]
10. Pilario, K.E.; Shafiee, M.; Cao, Y.; Lao, L.; Yang, S.-H. A Review of Kernel Methods for Feature Extraction in Nonlinear Process Monitoring. *Processes* **2020**, *8*, 24. [CrossRef]
11. Chang, Y.W.; Wang, Y.C.; Tao, L.; Wang, Z.J. Fault diagnosis of a mine hoist using PCA and SVM techniques. *J. China Univ. Min. Technol.* **2008**, *18*, 327–331. [CrossRef]
12. Li, C.-H.; Kuo, B.-C.; Lin, C.-T. LDA-Based Clustering Algorithm and Its Application to an Unsupervised Feature Extraction. *IEEE Trans. Fuzzy Syst.* **2011**, *19*, 152–163. [CrossRef]
13. Liu, C.; Wechsler, H. Independent component analysis of Gabor features for face recognition. *IEEE Trans. Neural. Netw.* **2003**, *14*, 919–928.
14. Nomikos, P.; MacGregor, J.F. Monitoring batch processes using multiway principal component analysis. *AIChE J.* **1994**, *40*, 1361–1375. [CrossRef]
15. Nomikos, P.; MacGregor, J.F. Multi-way partial least squares in monitoring batch processes. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 97–108. [CrossRef]
16. Ku, W.; Storer, R.H.; Georgakis, C. Disturbance detection and isolation by dynamic principal component analysis. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 179–196. [CrossRef]
17. Choi, S.W.; Lee, C.; Lee, J.-M.; Park, J.H.; Lee, I.-B. Fault detection and identification of nonlinear processes based on kernel PCA. *Chemom. Intell. Lab. Syst.* **2005**, *75*, 55–67. [CrossRef]
18. Cao, L.J.; Chua, K.S.; Chong, W.K.; Lee, H.P.; Gu, Q.M. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing* **2003**, *55*, 321–336. [CrossRef]
19. Joe Qin, S.; Dunia, R. Determining the Number of Principal Components for Best Reconstruction. *IFAC Proc. Vol.* **1998**, *31*, 357–362. [CrossRef]
20. MacGregor, J. Multivariate Statistical Approaches to Fault Detection and Isolation. *IFAC Proc. Vol.* **2003**, *36*, 549–554. [CrossRef]
21. Lee, J.-M.; Yoo, C.; Lee, I.-B. Statistical process monitoring with independent component analysis. *J. Process Control.* **2004**, *14*, 467–485. [CrossRef]
22. Chen, Y.; Zi, Y.; Cao, H.; He, Z.; Sun, H. A data-driven threshold for wavelet sliding window denoising in mechanical fault detection. *Sci. China Technol. Sci.* **2014**, *57*, 589–597. [CrossRef]
23. Yongguo, L. Feature Subset Selection Based on Genetic Algorithm. *Comput. Eng.* **2003**, *29*, 19–20.
24. ElAlami, M.E. A filter model for feature subset selection based on genetic algorithm. *Knowl. Based Syst.* **2009**, *22*, 356–362. [CrossRef]
25. Iba, K. Reactive power optimization by genetic algorithm. *IEEE Trans. Power Syst.* **1994**, *9*, 685–692. [CrossRef]
26. Lingaraj, H. A Study on Genetic Algorithm and its Applications. *Int. J. Comput. Sci. Eng.* **2016**, *4*, 139–143.
27. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [CrossRef]
28. Xiaohang, C.; Huifeng, X. Design of Elitist Adaptive Genetic Algorithm in Arrival Aircrafts Scheduling. *Comput. Commun.* **2006**, *24*, 91–94.
29. Liao, M. Study on the Effect of Cataclysm Operator on Genetic Algorithm. *Comput. Eng. Appl.* **2005**, *41*, 54–57.
30. Peng, Z. A Partheno-genetic Algorithm Based on Cataclysm. *J. Hubei Automot. Ind. Inst.* **2007**, *2*, 19–21.
31. Crammer, K.; Singer, Y. On the Learnability and Design of Output Codes for Multiclass Problems. *Mach. Learn.* **2002**, *47*, 201–233. [CrossRef]
32. Lyman, P.R.; Georgakis, C. Plant-wide control of the Tennessee Eastman problem. *Comput. Chem. Eng.* **1995**, *19*, 321–331. [CrossRef]

33. Wang, H.; Song, Z.; Hui, W. Statistical process monitoring using improved PCA with optimized sensor locations. *J. Process Control.* **2002**, *12*, 735–744. [CrossRef]
34. Wang, H.; Song, Z.; Ping, L.I. Improved PCA with application to process monitoring and fault diagnosis. *J. Chem. Ind. Eng.* **2001**, *52*, 471–475.
35. Wang, K.; Chen, J.; Song, Z. Performance Analysis of Dynamic PCA for Closed-Loop Process Monitoring and Its Improvement by Output Oversampling Scheme. *IEEE Trans. Control. Syst. Technol.* **2019**, *27*, 378–385. [CrossRef]