Industrial Process Fault Detection Based on Entropy Score Contribution Analysis

FEI LI1,2,3,4, XIAOQIANG LIU5, LIANGLIANG SHANG5, GUOZHU WANG6, ZIWEI PAN1, AND YINING SUN4

1 Post-Doctoral Scientific Research Workstation, Ma’anshan University, Ma’anshan, Anhui 243100, China
2 School of Electrical and Information Engineering, Anhui University of Technology, Ma’anshan, Anhui 243032, China
3 Anhui Province Key Laboratory of Special Heavy Load Robot, Anhui University of Technology, Ma’anshan, Anhui 243032, China
4 Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, Anhui 230031, China
5 School of Electrical Engineering, Nantong University, Nantong, Jiangsu 226019, China
6 Department of Automatic Control, Henan Institute of Technology, Xinxiang, Henan 453003, China

Corresponding author: Fei Li (lifei41@ahut.edu.cn)

This work was supported in part by the Youth Project of National Natural Science Foundation of China under Grant 61903003, in part by the Youth Project of Natural Science Foundation of Anhui Province under Grant 2008085QE227, in part by the Key Research Project of Natural Science in Colleges and Universities of Anhui Province under Grant KJ2019A0051, in part by the Open Project of Anhui Provincial Key Laboratory of Special Heavy-Duty Robots under Grant TZJQR001-2021, and in part by the Ma’anshan University Postdoctoral Research Launch Fund under Grant PRC2021002.

ABSTRACT Aiming at the problem of poor fault detection ability in complex industrial processes, this paper proposes a fault detection method of entropy score contribution analysis, which can better determine the number of samples with a high contribution rate and achieve good fault detection results. First, normalize the input data to obtain the Euclidean distance of each line vector; Secondly, the number of samples with a high contribution rate is determined by entropy score contribution and determine the initial threshold; Thirdly, a parameter adaptive strategy is proposed to set the threshold of statistics; Finally, fault detection is carried out on the statistical data, and the size of fault detection rate and false alarm rate is calculated. In the fourth part of this paper, we use this fault detection method to detect the faults from the Tennessee Eastman process. The simulation results show that the method proposed has more advantages than the traditional methods, and can be effectively applied to complex industrial processes.

INDEX TERMS Euclidean distance, entropy score contribution, fault detection, false alarm rate, fault detection rate.

I. INTRODUCTION

With the development and progress of society, the operation state of industrial processes often changes [1], and the whole industrial process continues to move towards intelligence, automation, integration, and so on. For the processes of chemical production, blast furnace ironmaking, and high-precision machining, if the fault detection and identification of the process cannot be carried out in time, it will cause major accidents.

In the past decades, multivariable process monitoring and data-driven methods have been widely used in fault detection and achieved great success in industrial process monitoring. For example, Jiang et al. [2] proposed the method of principal component analysis (PCA), Kano et al. [3] proposed the method of independent component analysis (ICA), Qi et al. [4] proposed the method of combining information entropy theory and principal component analysis (ECA) solves the problem of the comprehensive evaluation of unit operation state in fault diagnosis of thermal power plants, which has high reliability and effectiveness. Schölkopf et al. [5] proposed the method of nuclear principal component analysis (KPCA), Zhang et al. [6] proposed the method of nuclear independent component analysis (KICA) for fault detection of industrial processes, Zhu et al. [7] proposed a kernel entropy component analysis (KECA) method and
supervised learning algorithm is introduced to extract the geometric features of the data to improve the accuracy of fault diagnosis. Zhang et al. [8] proposed a decentralized process monitoring method based on multi-block kernel partial least squares (KPLS) to monitor large-scale nonlinear processes. Liu et al. [9] proposed a multi-level process monitoring method for the continuous annealing process based on PCA. Lou et al. [10] proposed a novel robust PCA scheme that achieves a high fault detection rate and low false alarm rate in the TE process simulation experiments. By considering the specificity of each block and the correlation between different blocks, Tong and others [11] developed an improved multi-block PCA method for the whole plant process. Downs and Vogel [12] proposed a method based on the combination of the cumulative sum model and principal component analysis and carried out fault detection for the Tennessee Eastman process. The simulation results show that this method has significant advantages over the traditional principal component analysis method.

Based on the idea of data dimensionality reduction, the above method mainly combines the principal component analysis method to process high-dimensional data and retain the most important features to improve the processing speed of data [13]. At the same time, it combines $T^2$ statistics and SPE statistics to monitor the whole industrial process. However, the above method applies to data with a strong correlation between variables. If the original data is not highly correlated, it will not achieve good dimensionality reduction. In addition, when the process data does not conform to the normal distribution, the value detected using $T^2$ statistics is unreliable, and the performance of fault detection is not satisfactory.

On the other hand, the use of data-driven methods for fault detection has also achieved some results [14]. The data-driven method analyzes a large number of process data and establishes a fault detection and diagnosis model. The main methods are partial least squares, multivariate statistical process monitoring, dynamic principal component analysis, neural networks, and so on [15], [16], [17], [18]. These methods have been widely used in practical complex industries [19]. Feng et al. [20] proposed a multimodal process fault detection strategy based on a standard k-nearest neighbor. This method is used to detect faults in the numerical simulation process and penicillin fermentation process. The results show that this method has a higher fault detection rate than principal component analysis and kernel principal component analysis. Liu et al. [21] proposed a multimodal process fault detection method based on a weighted distance neighborhood selection strategy. This method first weights the Euclidean distance reasonably and then selects the field of sample points according to the new weighted distance, which can effectively avoid the problem of incomplete data information. The simulation results of the TE process show that this method can achieve good results in the application of the density-based detection method.

Recently, Cai et al. [22] proposed a data-driven early fault diagnosis method for permanent magnet synchronous motor based on a Bayesian network. The method of wavelet threshold de-noising and minimum entropy deconvolution was used to improve the signal-to-noise ratio. The method of complementary set decomposition was used to extract signal eigenvalues, and the Bayesian network was used to identify faults. Experimental data verified the feasibility and effectiveness of this method. Kong et al. [23] proposed a fault diagnosis method with multiple modular redundant closed-loop feedback, and uses sensor data and system parameters to establish a dynamic Bayesian network for fault diagnosis. This method can dynamically evaluate the performance of the system. The experiment of a dual module redundant control system of a subsea blowout preventer shows that the proposed method has high fault detection accuracy. Kong et al. [24] proposed a sensor placement method for a hydraulic control systems based on a discrete particle swarm optimization algorithm for fault diagnosis of a hydraulic control systems, which is used to determine the number and placement position of sensors. In this paper, a typical multi-loop hydraulic control system is simulated, and the results show that the proposed method has a fast convergence effect, and the robustness of this method is better than that of traditional methods. Wang et al. [25] proposed a model named recursive hybrid variable monitoring (RHVM) to solve the problem of process monitoring and through the simulation example verify the feasibility of the method. Cai et al. [26] proposed fault detection and diagnostic method for diesel engines. This method combines rule-based algorithm and Bayesian networks (BNs) or Back Propagation neural networks (BPNNs) to study the influence of various interference factors on fault diagnosis and achieves good results.

For complex industrial processes, once faults occur, industrial processes will be shut down and paralyzed. Therefore, it is necessary to find an effective method to timely mine the change information in the process and monitor the whole industrial process. The main contributions of this paper are as follows: Firstly, the Euclidean distance of each line vector of test set data and training set data is obtained by processing the statistics; Secondly, the entropy value of Euclidean distance of training set data is calculated, and the comprehensive score of each sample is obtained through weighting operation; Thirdly, the number of samples with large contribution rate and the size of initial threshold are determined by the method of the cumulative contribution of entropy fraction; Then a parameter adaptive strategy is proposed to dynamically select the threshold size for the problem of poor fault detection effect; Finally, fault detection is carried out for TE process, and the fault detection rate and false alarm rate are calculated and compared with the traditional principal component analysis and entropy component analysis methods.

The specific structure of the article is as follows. In Section II, the preparatory work in the early stage of the article is introduced to prepare for the subsequent fault detection. Calculate the Euclidean distance of each row of
The calculation formula is as follows:

\[ \text{dimensional space is the actual distance between two points.} \]

\[ \text{Euclidean distance in two-dimensional and three-dimensional space.} \]

\[ \text{It can also be understood as the true distance between two points in multidimensional space.} \]

\[ \text{Mahalanobis distance is also a measure of distance, which} \]

\[ \text{can be seen as an update of Euclidean distance, and is an effective method} \]

**II. PREPARATION**

In complex industrial processes, the characteristics of variables will change randomly, resulting in a change in their values. Getting the information we need from the variable itself is a problem worth studying. Therefore, we need to deal with the data before the experiment. In this section, we mainly introduce the preparatory work of the article and propose a method based on entropy score contribution analysis.

**A. DATA PREPROCESSING**

Before analyzing the data, we usually have to deal with the data first. Data generally includes training sets and test sets, which contain a lot of information. The order of magnitude of each group of data may vary greatly, which makes it impossible for us to directly analyze features and laws.

Assume that the variables in the whole industrial process are \( X = (x_1, x_2, \ldots, x_m) \), among \( x_i \in \mathbb{R}^{(n \times 1)} \) (\( i = 1, 2, \ldots, m \)), and the variable \( x_i \) is independent of each other and follows Gaussian distribution. We have \( x_i \sim N (\mu_i, \sigma_i^2) \), among \( \mu_i \) means the average value and \( \sigma_i^2 \) represents variance. \( m \) represents the number of observed variables, and \( n \) represents the number of observation data. There are many methods of data standardization, we use the standard deviation standardization method, which is defined as:

\[ x'_i = \frac{(x_i - \mu_i)}{\sqrt{\sigma_i^2}} \]  

where \( x'_i \) refers to the data after standardization. Through data standardization, we can ensure that the dimensions of each indicator data are at the same level, facilitating subsequent calculations.

**B. EUCLIDEAN DISTANCE**

Euclidean distance is the most common distance measure, which measures the absolute distance between two points in multidimensional space, that is, the shortest straight-line distance between two points. It can also be understood as the true distance between two points in multidimensional space. Euclidean distance in two-dimensional and three-dimensional space is the actual distance between two points. The calculation formula is as follows:

\[ \text{dist}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]  

Mahalanobis distance is also a measure of distance, which can be seen as an update of Euclidean distance, and is an effective method to calculate the similarity between two unknown sample sets. For a multi-variable vector \( X = (x_1, x_2, \ldots, x_m)^T \) with a mean value of \( \mu = (\mu_1, \mu_2, \ldots, \mu_m)^T \) and the covariance matrix of \( \sigma \) its Mahalanobis distance is calculated as follows:

\[ D_M(x) = \sqrt{(x - \mu)^T \sigma^{-1} (x - \mu)} \]  

Mahalanobis distance exaggerates the effect of small variable changes. In addition, when the covariance matrix changes, the distance value will fluctuate greatly, and the instability is strong. In comparison, the Euclidean distance will be more stable. Therefore, this paper takes Euclidean distance as the research object.

**C. ENTROPY SCORE CONTRIBUTION**

In many data-driven methods in the past, determining the number of principal components is an important problem, and principal component analysis is usually used. The fault detection method obtains the eigenvalues by calculating the covariance and uses the idea of dimension reduction to determine the number of principal components and threshold to calculate the contribution. However, this method can not achieve good results for data with weak correlation. In this case, the entropy method can better determine the number of principal components and determine the threshold value.

In information theory, entropy is a measure of uncertainty. The entropy method is a method to determine the weight of indicators by calculating the information entropy of the observed values of each indicator and according to the impact of the relative change degree of each indicator on the overall system. The method selects the number of principal components according to the information entropy value to retain the information of the original data as much as possible, which is widely used in the field of fault detection. It is worth noting that by processing the data and calculating the Euclidean distance of each line of vector, we can convert the high-latitude data into the low-latitude data. Therefore, the article does not need to introduce a kernel function to complicate the calculation.

In this paper, a fault detection method based on entropy score contribution analysis is proposed. By calculating the comprehensive score of each sample and using the cumulative contribution method, the number of samples whose Euclidean distance between each line of vectors in the training set data has a high contribution rate is determined and judge the size of the initial threshold. See algorithm steps for specific operation steps. Select \( m \) indicators and \( n \) samples, then \( X_{ij} \) represents the value of the \( j \) index of the \( i \) sample, where \( i = 1, 2, \ldots, n, j = 1, 2, \ldots, m \).

Calculate the proportion of the \( i \)th sample in the \( j \)th index:

\[ P_{ij} = \frac{X_{ij}}{\sum_{i=1}^{n} X_{ij}} \]
Calculate the entropy value of the \( j \)-th index:

\[
E_j = -K \sum_{i=1}^{n} (P_{ij} \times \log(P_{ij}))
\]  

(5)

The parameter \( K \) should be given in advance as a constant, and its calculation formula is:

\[
K = \frac{1}{\log(n)}
\]  

(6)

among, \( n \) is the number of samples. Calculate the difference coefficient of index \( j \) and its value directly affects the weight.

\[
D_j = 1 - E_j
\]  

(7)

Calculate the weight of the evaluation index:

\[
W_j = \frac{D_j}{\sum_{j=1}^{m} D_j}
\]  

(8)

Calculate the comprehensive score of each sample:

\[
S_j = \sum_{j=1}^{m} W_j \times X_{ij}
\]  

(9)

Through the above algorithm steps, we propose a method of entropy score contribution. First, calculate the entropy value of the sample and get the comprehensive score of each sample, and then determine the number of samples with large contribution rate by calculating the cumulative contribution degree. Because the data type of Euclidean distance is simple, we set the threshold corresponding to the number of samples with high contribution rate as the initial threshold. The specific method will be described in detail in Section III.

III. ALGORITHM

A. CALCULATION OF EUCLIDEAN DISTANCE

In this section, we mainly introduce the generation program of Euclidean distance between each row of vectors in training data, and its pseudo code is shown in algorithm 1.

Through algorithm 1, we can get the Euclidean distance of vectors in each row of the training data. Similarly, we use the same method to calculate the vector distance of each row of the test data and record the value as \( \text{dist}_2 \). Through the calculation of algorithm 1, we can get that the data type of Euclidean distance statistic is a simple column vector, and this statistic is also the object of follow-up research in this article.

B. DETERMINE THE INITIAL THRESHOLD OF STATISTIC

Entropy is an important tool used to describe the uncertainty of random variables in the field of data-driven [27]. In short, entropy reflects the chaos of a system. The more chaotic a system is, the larger its entropy is; The more orderly, the smaller the entropy. Therefore, entropy is an important statistical index that can be used to describe whether the process state is normal or not. The traditional entropy method generally only calculates the entropy value and the weight of each index, which is highly dependent on the sample. With the change in the modeling sample, the weight will also change. Selecting the number of principal components by this method does not have high reliability. This paper proposes an entropy score contribution analysis method for fault detection. By modifying the index weight and calculating the comprehensive score of each sample. Then, the number of samples with a high contribution rate is determined by calculating the cumulative contribution degree from the entropy comprehensive score of samples.

Through the calculation of algorithm 1, we can obtain the Euclidean distance between the vectors in each row of the training set sample, and further determine that the data type of the Euclidean distance statistic is a column vector with the number of rows as the number of samples and the number of columns as 1. The distance can directly reflect the discrete relationship between various data, and can effectively reduce the impact of the error. The data type to calculate the distance value is simple, so the Euclidean distance becomes the research object of this paper. We use the entropy score contribution analysis method to determine the number of samples with a large contribution rate. At the same time, the size of the initial threshold is the number of samples with a large contribution rate in \( \text{dist}_1 \) data. The pseudo-code of the algorithm is shown in algorithm 2.

C. PARAMETER ADAPTIVE STRATEGY

Assume that the multivariate in the whole industrial process is \( X = (x_1, x_2, \ldots, x_m) \in R^{n \times m} \) and the variable \( x_i \) is independent of each other and follows Gaussian distribution. When the variable obeys the normal distribution, it can be calculated according to 3\( \sigma \) principles (\( \sigma \) means standard deviation) to judge the threshold. Its expression is as follows:

\[
m = E(x_i) + \mu \times \text{std}(x_i)
\]  

(10)

Algorithm 1 Generation of Euclidean Distance

Require: The training data \( X_{\text{train}} \)

Ensure: Euclidean distance between vectors in each row \( \text{dist}_1 \)

1: Process the training set data and calculate the mean and standard deviation;
2: The standard deviation normalization method is used to process training data;
3: \([X_{\text{row}}, X_{\text{col}}] = \text{size}(X_{\text{train}});
4: \text{dist}_1 = [ ];
5: \mu = \text{mean}(X_{\text{train}});
6: for \( i = 1 \) to \( X_{\text{row}} \) do
7: \( y_i = X_{\text{train}}(i,:);
8: \text{dist} = (\sum(y_i - \mu)^2)^{\frac{1}{2}};
9: \text{dist}_1 = [\text{dist}];
10: \text{end for}
11: A = \text{dist}_1;
Algorithm 2 Determine the Number of Samples With High Contribution Rate and Initial Threshold

Require: Euclidean distance between vectors in each row of the test data dist
Ensure: The number of samples with high contribution rate m and the initial threshold Tu

1. \( x = \text{dist}_1 \);
2. \([n, m] = \text{size}(x)\);
3. \( \text{lamda} = \text{ones}(n, m)\);
4. \( k = -\log(n)\);
5. \textbf{for} \( i = 1 \) to \( n \) \textbf{do}
6. \( p(i, 1) = \frac{x(i, 1)}{\text{sum}(x(:, 1))}\);
7. \( e(i, 1) = p(i, 1) \times \log(p(i, 1))\);
8. \textbf{end for}
9. \( E = -k \times \text{sum}(e(:, 1))\);
10. \( d = 1 - E\);
11. \( w = \frac{\text{sum}(d)}{w \times \text{lamda}(1)}\);
12. \( w = \frac{\text{sum}(w \times \text{lamda})}{\text{sum}(w \times x(i, 1))}\);
13. \textbf{for} \( i = 1 \) to \( n \) \textbf{do}
14. \( x(i, 1) = \text{sum}(w \times x(i, 1))\);
15. \textbf{end for}
16. \( m = 1\);
17. \textbf{while} \( \text{sum}(s(1:n)) / \text{sum}(s) < 0.95 \) \textbf{do}
18. \( n = m + 1\);
19. \textbf{end while}
20. \( Tu = \text{dist}_1(n)\);

Among them, \( \mu \) is a coefficient greater than zero, \( E(x_i) \) is the mathematical expectation of \( x_i \), \( \text{std}(x_i) \) is the standard deviation of \( x_i \), \( m \) is the threshold. Data in industrial processes have non-negative characteristics, so we can determine the value of \( \mu \) as 3 according to the principle of the confidence interval. Therefore, equation (10) can be written as:

\[ m = E(x_i) + 3 \times \text{std}(x_i) \]  \hspace{1cm} (11)

However, in a complex industrial process, variables will change randomly, such as step, vibration, constant change and other characteristics. The existence of these factors may cause variables to not conform to the law of normal distribution. Therefore, the threshold value calculated by equation (10) cannot be applied to all cases. At the same time, the initial threshold obtained by determining the number of samples with a large contribution rate in algorithm 2 can also play a good role in fault detection. From the analysis of equation (10), we can see that the data type of threshold \( m \) is the same as the Euclidean distance calculated from the vectors of each line of training data. We can determine the threshold value of equation (10) as \( M \) according to the same method as the initial threshold value, and introduce a parameter called influence factor \( K \). We optimize the threshold size through adaptive strategy, and dynamically select the initial threshold value \( Tu \) and threshold value \( M \) through the influence factor \( K \). After the final threshold is determined, the TE process can be detected, and good detection results can be obtained.

The TE process is simulated and analyzed to verify the feasibility of parameter adaptive adjustment. We choose Fault 1 as the reference object, which is only given as an example. The detection result of Euclidean distance statistic is shown in Figure 1. We can see that when the adaptive adjusted threshold is selected as the control limit of fault detection, the detection rate and false alarm rate are more advantageous than the initial threshold. The pseudo-code of the parameter adaptive adjustment strategy is shown in algorithm 3.

D. FAULT DETECTION STEPS

In this part, we will introduce the fault detection steps based on the entropy score contribution analysis method. Its specific steps are as follows:

1. Calculated the mean and standard deviation of the training data and the Euclidean distance of each vector of the training data is calculated according to algorithm 1;
2. According to algorithm 2, calculate the entropy value of the sample and the comprehensive score of each sample, analyze the entropy value score, and determine the number of samples with high contribution rate and the size of the initial threshold \( Tu \) according to the cumulative contribution;
3. According to the principle of a confidence interval and the determined number of samples with high contribution rate, we can determine the value of the second threshold \( M \);
4. The threshold is optimized by the parameter based adaptive adjustment strategy of algorithm 3 to prepare for subsequent fault detection;
(5) Compare the Euclidean distance calculated in the test data with the threshold $T_U$ value and draw a conclusion whether there is a fault.

IV. EXPERIMENTAL SIMULATION AND RESULT ANALYSIS

In this section, we apply the fault detection method based on the contribution of entropy score to the TE process to detect the size of data false alarm rate and fault detection rate in the industrial process.

A. INTRODUCTION TO TE PROCESS

According to the actual chemical reaction process, Eastman company of the United States developed a chemical model simulation platform, namely the Tennessee Eastman process. This platform is an open and very challenging chemical process test platform [28], [29]. The process flow diagram of the TE process is shown in Figure 3.

A total of 52 measured values can be collected in the TE process, including 41 process variables and 11 process variables. In this process, a set of 21 faults can be simulated and used to evaluate the performance of the monitoring method. Among the 21 faults provided, the main feature of faults 1-7 is the step change of process variables; faults 8-12 are caused by random changes in process variables [30]; fault 13 is caused by the slow change of reaction kinetics; faults 14 and 15 are caused by valve sticking; The type of faults 16-20 is unknown; fault 21 is caused by the valve of feed channel 4 being fixed in a constant position [31].

B. DATA PREPARATION AND PARAMETER DETERMINATION

The data set of the TE process is divided into two types: training sets and test sets. Each training set or test set represents a fault. The data in the TE data set is obtained from 22 different simulation runs and each sample has 52 observation variables. The training sample without fault is obtained under a 25h operation simulation, and the total number of observation data is 500. Training data is used to complete standardization processing and other work. It should be noted that the test set sample with faults was obtained under 48h running simulation. The faults were introduced at 8h, and each test data contains 960 samples, and all faults are introduced from the 161st sample to the end [32].

C. ANALYSIS OF SIMULATION RESULTS

In order to compare the fault detection ability of the method proposed in this paper with that of principal component analysis and entropy component analysis, in this section, we use fault 4 with step fault characteristics, fault 12 with random change characteristics, fault 14 with viscous change characteristics and faults with unknown change characteristics 19 are taken as the test data of the article. There are many similar types of faults in the whole TE process, and the four types of faults proposed are only shown as examples. The monitoring results are shown in Figures 4-7. Among them, sub-graph (a) is the fault detection result of Euclidean distance statistic based on the entropy score contribution analysis in the article, sub-graphs (b), (c), and sub-graphs (d), (e) are the detection results of $T^2$ statistics and $SPE$ statistics.
in principal component analysis and entropy component analysis, respectively.

The fault detection results of fault 4 with step change characteristics are shown in Figure 4, compared with the $T^2$ statistic detection in the traditional principal component analysis and entropy component analysis methods, the method proposed in this paper has a higher fault detection rate (indicates the probability of alarm in case of failure) and a similar false alarm rate (indicates the probability of alarm when no fault occurs). At the same time, in the detection of fault 4, compared with the $SPE$ statistics detection of the other two methods, the method proposed in this paper has more advantages in terms of product false alarm rate.

In order to detect the performance under the condition that the fault characteristics are random changes, fault 12 is used as the test data, and the test results are shown in Figure 5. The results show that the fault detection rate of the method proposed in this paper is slightly lower than that of the other two methods, and the results are similar and within our acceptable range. However, in terms of data false alarm rate, the method in this paper has more advantages in terms of product false alarm rate.

To detect the performance of the fault detection method under the condition that the fault characteristic is a viscous change caused by valve viscosity, fault 14 is used as the test data, and the test results are shown in Figure 6. Comparing the Euclidean distance statistic detected by the method proposed in the article with the $T^2$ statistic and $SPE$ statistic detected by the principal component analysis and entropy component analysis, the three statistics show good results in the fault detection rate of data, but the method proposed in the article has more advantages in the false alarm rate of data.

The detection of unknown changes in statistics is shown in Figure 7. We selected fault 19 as the test data for detection. From the figure, we can see that the detection effect of each statistic is not good, whether it is the data fault detection rate or the data false alarm rate. In the TE process, there are several groups of data similar to the data type of fault 19. These data are complex, which brings us great trouble. This is because the data of the TE process has time-varying, complex, and nonlinear characteristics, and the process is mixed with the influence of noise, which leads to an unsatisfactory detection effect.

In the industrial process, the two factors of product fault detection rate and false alarm rate are generally considered in the fault detection of the whole process. Detect 21 faults in the TE process and the comparison results of fault detection rate and false alarm rate are shown in Table 1. Among them, through comprehensive consideration, we show the statistical

![FIGURE 3. TE process flow chart.](image-url)
FIGURE 4. Statistical results of TE process fault 4.
(a) Detection of Euclidean distance statistic by ESCA method
(b) Detection of $T^2$ statistic in principal component analysis
(c) Detection of $SPE$ statistic in principal component analysis
(d) Detection of $T^2$ statistic in entropy component analysis
(e) Detection of $SPE$ statistic in entropy component analysis

FIGURE 5. Statistical results of TE process fault 12.
(a) Detection of Euclidean distance statistic by ESCA method
(b) Detection of $T^2$ statistic in principal component analysis
(c) Detection of $SPE$ statistic in principal component analysis
(d) Detection of $T^2$ statistic in entropy component analysis
(e) Detection of $SPE$ statistic in entropy component analysis
FIGURE 6. Statistical results of TE process fault 14.

(a) Detection of Euclidean distance statistic by ESCA method
(b) Detection of $T^2$ statistic in principal component analysis
(c) Detection of $SPE$ statistic in principal component analysis

(d) Detection of $T^2$ statistic in entropy component analysis
(e) Detection of $SPE$ statistic in entropy component analysis

FIGURE 7. Statistical results of TE process fault 19.

(a) Detection of Euclidean distance statistic by ESCA method
(b) Detection of $T^2$ statistic in principal component analysis
(c) Detection of $SPE$ statistic in principal component analysis

(d) Detection of $T^2$ statistic in entropy component analysis
(e) Detection of $SPE$ statistic in entropy component analysis
As shown in the table, compared with the traditional principal component analysis and entropy component analysis method, the fault detection method based on the contribution of entropy score in the simulation analysis of 21 faults in the TE process shows that the product fault detection rate and false alarm rate show good results. In the operation of industrial systems, unknown faults are inevitable [33], when Euclidean distance statistics are used to monitor industrial process variables, it is insensitive to some faults (such as fault 3, fault 9, fault 19, and fault 21), which is reflected in the effect of fault false alarm rate. The reason is that the correlation of TE process variables is nonlinear, and there are many noises in the variables, which will make the information saved by the statistics not obvious, so it will lead to the detection effect of some faults is not significant. In the analysis of 21 faults, most of the data results show superiority, so the fault detection method proposed in this paper can be applied to some complex industrial processes.

V. CONCLUSION

In order to improve the performance of industrial process fault detection methods, this paper proposes a fault detection method based on entropy score contribution analysis. This method directly starts from the variable itself and calculates the Euclidean distance between the vectors in each row as the statistics to be measured. Its data type is simple, it can intuitively observe the fault information between statistical variables, reduce the error between information, and has high reliability. In addition, we calculate the entropy value of the statistics and obtain the comprehensive score of each sample by modifying the index weight. We adopt the cumulative summation method for the comprehensive scores of each sample.
sample and obtain the number of samples with a high contribution rates and determine the size of the initial threshold. Finally, we propose a parameter adaptive strategy to dynamically select threshold size which greatly reduces the impact of errors to obtain good fault detection results.

The fault detection method proposed in this paper is applied to the simulation analysis of the TE process. Compared with the traditional principal component analysis and entropy component analysis methods, the product fault detection rate and false alarm rate of the method proposed in this paper can show a good detection effect in multiple groups of test data, reaching an acceptable level, so that this method can be better applied to complex industrial processes. In the next work, we will also consider that when there are delays in industrial processes and the presence of noise in industrial processes, the existence of factors will lead to more complex fault detection problems and bring us greater challenges [34].

REFERENCES

[1] L. Shang, Z. Yan, J. Li, A. Qiu, and H. Zhang, “Canonical residual based incipient fault detection method for industrial process,” in Proc. Chin. Control Decis. Conf. (CCDC), 2020, pp. 987–992.

[2] Q. Jiang, X. Yan, and B. Huang, “Performance-driven distributed PCA process monitoring based on fault-variant relevant variable selection and Bayesian inference,” IEEE Trans. Ind. Electron., vol. 63, no. 1, pp. 377–386, Jan. 2016.

[3] M. Kano, S. Tanaka, S. Hasebe, I. Hashimoto, and H. Ohno, “Combined multivariate statistical process control,” IFAC Proc. Volumes, vol. 37, no. 1, pp. 281–286, Jan. 2004.

[4] M. Qi, Z. Fu, Y. Jing, and Y. Ma, “Comprehensive evaluation method of thermal power unit based on information entropy and principal component analysis,” Chin. J. Electr. Eng., vol. 33, no. 2, pp. 58–64, 2013.

[5] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” Neural Comput., vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

[6] Y. Zhang, J. An, and H. Zhang, “Monitoring of time-varying processes using kernel independent component analysis,” Chem. Eng. Sci., vol. 88, pp. 23–32, Jan. 2013.

[7] Z. Zhu, D. Wu, and Y. Yue, “Engine wear fault diagnosis based on supervised kernel entropy component analysis,” J. Syst. Simul., vol. 34, no. 1, pp. 45–52, 2022.

[8] Y. Zhang, H. Zhou, S. J. Qin, and T. Chai, “Decentralized fault diagnosis of large-scale processes using multiblock kernel partial least squares,” IEEE Trans. Ind. Informat., vol. 6, no. 1, pp. 3–10, Feb. 2010.

[9] Q. Liu, S. J. Qin, and T. Chai, “Decentralized fault diagnosis of continuous annealing processes based on multilevel PCA,” IEEE Trans. Autom. Sci. Eng., vol. 10, no. 3, pp. 687–698, Jul. 2013.

[10] Z. Lou, Y. Wang, S. Lu, and P. Sun, “Process monitoring using a novel robust PCA scheme,” Ind. Eng. Chem. Res., vol. 60, no. 11, pp. 4397–4408, Mar. 2021.

[11] C. D. Tong and X. F. Yan, “A novel decentralized process monitoring scheme using a modified multiblock PCA algorithm,” IEEE Trans. Autom. Sci. Eng., vol. 14, no. 2, pp. 1129–1138, Apr. 2017.

[12] J. J. Downs and E. F. Vogel, “A plant-wide industrial process control problem,” Comput. Chem. Eng., vol. 17, no. 3, pp. 245–255, Mar. 1993.

[13] L. Shang, A. Qiu, P. Xu, and F. Yu, “Canonical variate nonlinear principal component analysis for monitoring nonlinear dynamic processes,” J. Chem. Eng. Jpn., vol. 55, no. 1, pp. 29–37, 2022.

[14] L. Shang, K. Shi, C. Ma, A. Qiu, and L. Hua, “Fault detection and identification based on explicit polynomial mapping and combined statistic in nonlinear dynamic processes,” IEEE Access, vol. 9, pp. 149050–149066, 2021.

[15] Y. Wang, Q. Jiang, and J. Fu, “Data-driven optimized distributed dynamic PCA for efficient monitoring of large-scale dynamic processes,” IEEE Access, vol. 5, pp. 18325–18333, 2017.

[16] G. Wang and S. Yin, “Quality-related fault detection approach based on orthogonal signal correction and modified PLS,” IEEE Trans. Ind. Informat., vol. 11, no. 2, pp. 398–404, Apr. 2015.
XIAOQIANG LIU received the graduate degree from the School of Electrical and Information Engineering, Anhui University of Technology, in 2021, where he is currently pursuing the master’s degree in control engineering. His research interest includes fault detection.

LIANGLIANG SHANG received the Ph.D. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2016. He is currently a Professor with the School of Electrical Engineering, Nantong University. His research interests include complex industrial process modeling, fault detection, and diagnosis.

GUOZHU WANG received the doctorate degree in control theory and control engineering from Northeast University, in 2017. He is an Associate Professor with the Henan Institute of Technology. His research interests include complex industrial process monitoring and fault diagnosis.

ZIWEI PAN received the graduate degree from the School of Mechanical Engineering, Beijing University of Science and Technology. He is the Dean of the Smart Manufacturing Engineering School, Maanshan University. His research interests include the operation performance and status of large mechanical equipment and equipment fault diagnosis technology.

YINING SUN received the graduate degree in precision instruments from the University of Science and Technology of China, in 1983. He is a Researcher and the Doctoral Supervisor of the Hefei Institutes of Physical Science, Chinese Academy of Sciences.