processes

MDPI

Article

Identification of Unknown Abnormal Conditions in Catalytic Cracking Process Based on Two-Step Clustering Analysis and Signed Directed Graph

Juan Hong [1], Jian Qu [1], Wende Tian [1,*], Zhe Cui [1], Zijian Liu [1], Yang Lin [2] and Chuankun Li [2]

[1] College of Chemical Engineering, Qingdao University of Science & Technology, Qingdao 266042, China; 0020010003@mails.qust.edu.cn (J.H.); QJ1143175623@163.com (J.Q.); cuizhequst@126.com (Z.C.); l707989840@163.com (Z.L.)

[2] State Key Laboratory of Safety and Control for Chemicals, SINOPEC Qingdao Research Institute of Safety Engineering, Qingdao 266071, China; linyang.qday@sinopec.com (Y.L.); lick.qday@sinopec.com (C.L.)

* Correspondence: tianwd@qust.edu.cn

Abstract: There are many unknown abnormal working conditions in industrial production. It is difficult to identify unknown abnormal working conditions because there are few relative sample and experience in this field. To solve this problem, a new identification method combining two-step clustering analysis and signed directed graph (TSCA-SDG) is proposed. Firstly, through correlation analysis and R-type clustering analysis, the variables are effectively selected and extracted. Then, a two-step clustering analysis was carried out on the selected variables to obtain the cluster results. Through the establishment of the signed directed graph (SDG) model, the causes of abnormal working conditions and their mutual influence are deduced from the mechanism. The application of the TSCA-SDG method in the catalytic cracking process shows that this method has good performance for abnormal condition identification.

Keywords: two-step clustering analysis; signed directed graph; catalytic cracking process; abnormal identification

Citation: Hong, J.; Qu, J.; Tian, W.; Cui, Z.; Liu, Z.; Lin, Y.; Li, C. Identification of Unknown Abnormal Conditions in Catalytic Cracking Process Based on Two-Step Clustering Analysis and Signed Directed Graph. *Processes* **2021**, *9*, 2055. https://doi.org/10.3390/pr9112055

Academic Editor: Andrea Petrella

Received: 28 October 2021
Accepted: 15 November 2021
Published: 16 November 2021

1. Introduction

As heavy and inferior crude oil becomes more and more popular, the fluidic catalytic cracking (FCC) process, as one of the core processes in the light-weight processing of heavy oil, has received more and more attention [1–3]. In China, the diesel and gasoline produced by FCC units account for about 30% and 70% of the finished diesel and gasoline [4]. With the application of distributed control system (DCS), the control of FCC has been computerized. At the same time, with the advent of big data era, the digitization and intelligence of FCC process have been developed widely.

The FCC process has some flammable, explosive chemicals and high temperature, high pressure conditions. The occurrence of accidents will cause serious casualties and property losses, as well as irreversible environmental pollution. With the development of computers, the industrial production process has become increasingly more automated, where abnormal alarms are mostly handled by operators. Due to the lack of ability and experience of operators, it is difficult to make correct judgments and take action quickly in case of abnormal occurrence, which may cause more serious subsequent accidents. According to industry statistics, abnormal events caused by operators accounted for about 70% of overall events [5]. Therefore, the current industrial production needs to introduce more effective computer system-based program for fault detection and diagnosis. Fault diagnosis technology has developed rapidly since the 1980s [6–8], which is generally divided into knowledge-based, mechanism-based and data-based technologies [9]. Data-based fault diagnosis technology does not have over reliance on rich expert experience and accurate

analytical models because it makes full use of the large amount of data generated during the operation of machinery and equipment. With the rapid development of industrial big data and computer technology, data-based fault diagnosis technology is more and more widely applied [10–13]. The DCS system also provides vitality to the application and innovation of data-driven methods in a chemical process failure study [14].

The data-based fault diagnosis technology can be classified as qualitative and quantitative methods, while the latter one can be further classified into two categories: statistical and non-statistical [15–17]. Cluster analysis belongs to the statistical technology, which is a typical unsupervised learning technology in the field of data mining and machine learning. Cluster analysis techniques can be used to explore and discover hidden patterns in data. The main division basis of cluster analysis method is the similarity relationship between the sample points, which is an autonomous division of the data sample set. In the clustering process, all the sample points in the same set are divided into several clusters, where the similarity of sample points in the same cluster structure is kept as high as possible but the similarity between different clusters is kept as low as possible. At present, the commonly used clustering algorithms include Two-Step Clustering (TSC) [18], K-means [19], Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [20], Gaussian Mixture Clustering (GMC) [21], Hierarchical Agglomerative Clustering (HAC) [22] and so on. DBSCAN can cluster dense datasets of any shape without unbiased results. However, when the density of the sample set is uneven or the clustering distance is very different, the clustering quality is poor. GMC can obtain elliptical clusters rather than circular ones with the mean and standard deviation. HAC is a bottom-up clustering algorithm; the disadvantage is that the computational complexity is too high and the efficiency is low. K-means has the advantages of simplicity, high efficiency, short time and low space complexity for large datasets. However, when the dataset is large, the result is prone to the local optimum. Moreover, K-means needs to set the value of K in advance and is, therefore, very sensitive to the selection of the K value [23]. TSC is a clustering method recently developed. It occupies fewer memory resources and has a fast computing speed for large datasets. TSC has an excellent clustering effect, so it is widely used in medical, nuclear engineering and other fields. In the identification of working conditions of industrial big data, TSC can accurately identify and cluster data of abnormal working condition. Although the above cluster methods have received in-depth development, their analysis of specific industrial mechanism is insufficient. Signed directed graph (SDG) is one of the labeling methods for mechanism analysis.

SDG is a qualitative fault identification method, which has the advantages of simple modeling and flexible reasoning. SDG is a good way to show the relationship between complex system variables and reveal the propagation path of potential hazards and failures. SDG has a wide range of applications and development. Yang et al. summarized the background and development of the SDG method, and reviewed three modeling methods of SDG and their application in the field of safety evaluation and fault diagnosis [24]. Gao et al. proposed a semi-quantitative validation method for a simulation model based on SDG and qualitative trends, where qualitative trends were added to the SDG model and the complete testing cases were produced by positive inference. The semiquantitative validation was carried out by comparing the testing cases with outputs of the simulation model in different scales [25]. Wu et al. determined candidate faults based on SDG backward inference from the alarm parameters. According to the candidate faults, SDG forward inference was applied to obtain candidate parameters and then identify real faults [26]. Guo et al. proposed a general framework for the translation of multi-attribute graphs. In order to discover and preserve the consistency of the generated nodes and edges, a spectral graph regularization based on a non-parametric Laplacian graph was designed [27].

This paper proposes a method for identifying unknown abnormal conditions in the catalytic cracking process by combining the two-step clustering analysis with the SDG model (TSCA-SDG). The TSCA-SDG method identifies abnormal working conditions

through two-step clustering analysis, and analyzes the propagation path of abnormal working conditions by the SDG model. The outline of this paper is organized as follows. In Section 2, the framework of the TSCA-SDG method is introduced in detail, followed by the principles of two-step clustering analysis and SDG. The excellent performance of TSCA-SDG is proved by a case study in Section 3. Section 4 provides a summary of this paper.

2. Proposed Method

The TSCA-SDG method consists of four parts: (1) data preprocessing; (2) feature extraction and selection; (3) two-step clustering analysis; (4) SDG modeling. Its framework is shown in Figure 1.

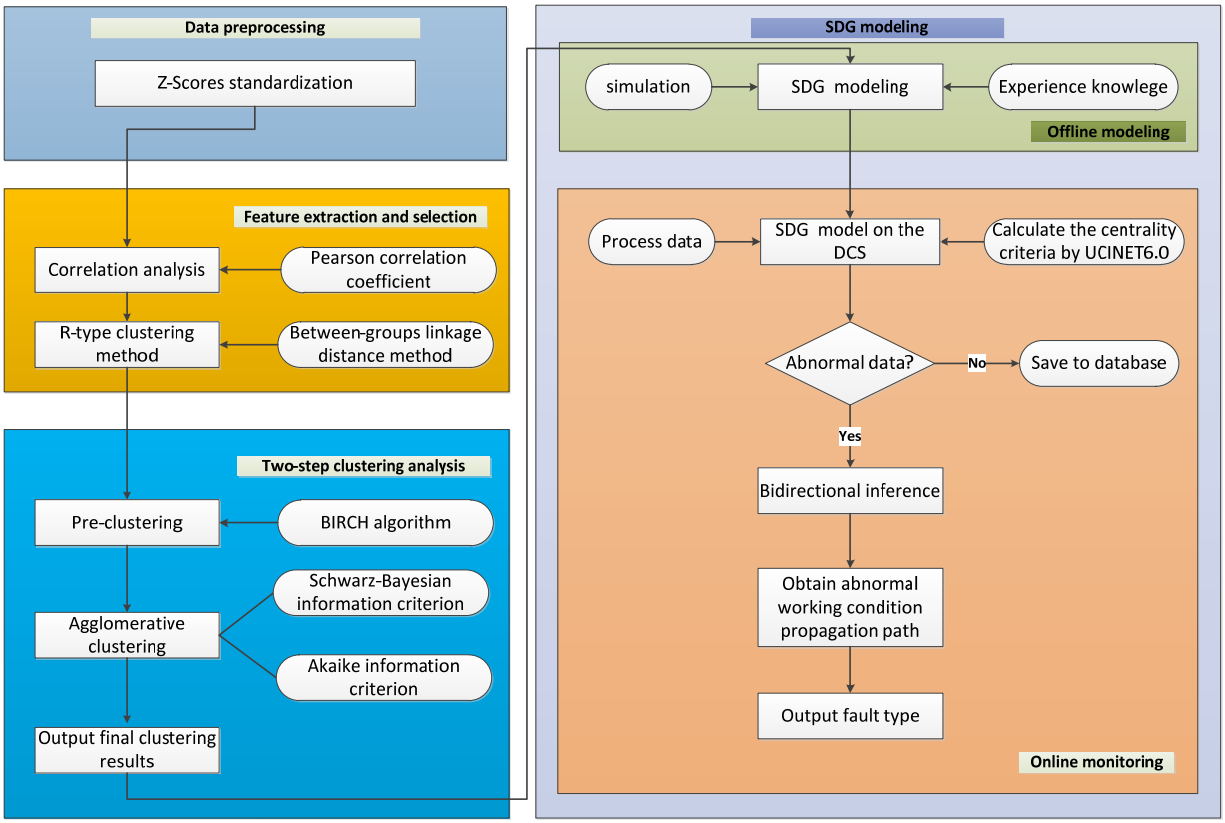

Figure 1. The framework of the TSCA-SDG method.

In the data processing part, Z-Score standardization is performed on the data of the control parameter and related variables to obtain data that remove the influence of magnitudes.

In the feature selection and extraction part, correlation analysis and R-type cluster analysis are carried out on the preprocessed variables. By calculating the Pearson correlation coefficient and the distance between variables, variables are selected and extracted to effectively achieve the purpose of dimensionality reduction.

In the two-step clustering analysis part, the variables after screening are clustered using the two-step clustering method. The optimal number of clusters is obtained through the Schwarz Bayesian Information Criterion for quick and effective clustering.

In the SDG model part, the SDG model is connected to the DCS system to realize process monitoring. For abnormal working condition data, the SDG model can accurately describe the fault characteristics as a consistent path through bidirectional inference. The abnormal type marked with characteristics can be output and displayed to the operator for the warning of potential abnormal occurrence.

2.1. Data Preprocessing

The dimensions of variables are different and their magnitudes vary greatly. For comparison of these data together, the data are preprocessed first. Z Scores standard deviation is used to eliminate the influence of dimension, where the mean value of the transformed data is 0 and the standard deviation is 1, as shown in Equation (1):

$$x_{ij}^{*} = \begin{cases} \frac{x_{ij}-\overline{x_j}}{S_j} & S_j \neq 0 \\ 0 & S_j = 0 \end{cases} \quad \begin{pmatrix} i = 1, 2, \cdots, n \\ j = 1, 2, \cdots, n \end{pmatrix} \tag{1}$$

where $\overline{x}_j$ is the mean of the data and S_j is the standard deviation of the data.

2.2. Feature Extraction and Selection

In the application of actual industrial big data, some closely related variables in industrial production show low correlation due to time lag and other reasons. If only the correlation analysis is considered, some variables that are correlated in practice may be ignored. Combining expert experience, this paper proposes a feature extraction and selection method that comprehensively considers correlation analysis and R-type clustering analysis to effectively solve this problem.

2.2.1. Correlation Analysis

For the relationship between variables, it is easy to think of the deterministic relationship between variables. Its characteristic is that when the value of one variable is determined, the value of other variables is also completely determined. Different from the deterministic relationship, there is an indeterminate relationship between variables. Its characteristic is that after a variable value is given, the value of another variable can change within a certain range. This non-deterministic relationship is called correlation. It must be studied with the help of statistical methods, which is also called statistical correlation [28].

The Pearson correlation coefficient is used to analyze the correlation of variables, as shown in Equation (2):

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2}$$

where n is the sample size and x_i and y_i are the variable values of the two variables, respectively.

2.2.2. R-Type Clustering Method

The R-type clustering method separates variables with large differences and clusters similar variables together. A few representative variables can be selected from similar variables to participate in other analyses to achieve the purpose of reducing the number of variables and dimensionality of variables.

The R-type clustering method used in this paper adopts agglomeration method. The process of agglomerative clustering is as follows. First, each observed individual is divided into a class. Then, the degree of closeness between all individuals is measured by the between-groups linkage distance method, and the closet individuals are grouped into a small class to form $n - 1$ classes. Next, the degree of closeness between the remaining observed individuals and subclasses is measured again, and the current closest individuals and subclasses are grouped into one class. The above process repeats until all the individuals are grouped together to form the largest group [29]. The flowchart of the R-type clustering method is shown in Figure 2.

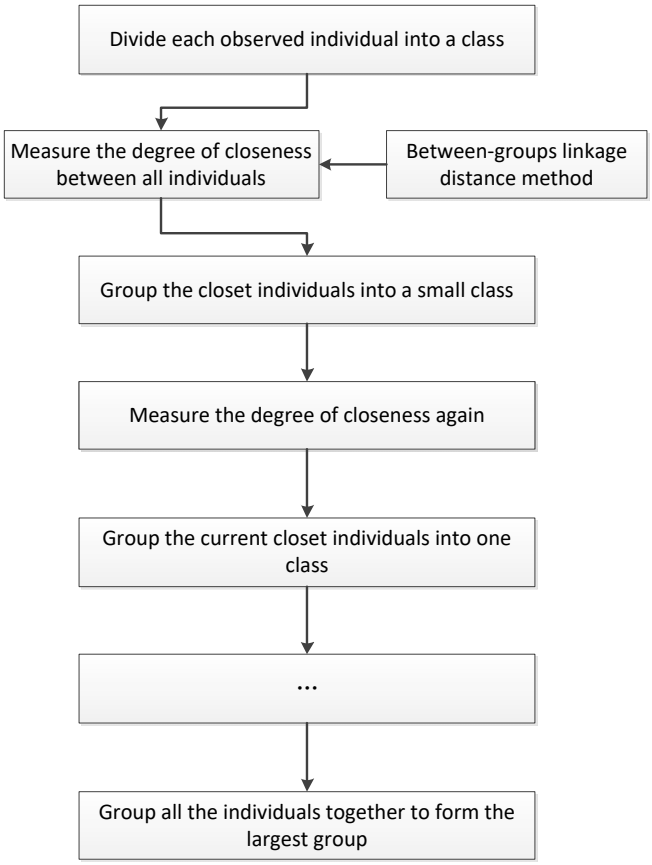

Figure 2. R-type clustering method flowchart.

The between-groups linkage distance is the average distance between an individual and each individual in the subclass. The between-groups linkage distance method overcomes the weakness that the nearest neighbor distance or the farthest neighbor distance is easily affected by extreme values as it uses the information of all distances between individuals and subclasses. During the agglomerative clustering, as the clustering progresses, the degree of closeness within the cluster gradually decreases. For n observed individuals, they can be agglomerated into a large class through $n - 1$ steps.

2.3. TSCA Method

Cluster analysis is an important part of the data mining discipline. It finds meaningful clusters from huge, seemingly chaotic data by mining the hidden patterns behind the data. The clustering algorithm is an unsupervised algorithm because there is no need to define the class in advance. Without taking the known classification information into consideration, all classification information can be generated by the clustering algorithm.

The two-step clustering algorithm is also called the two-stage clustering algorithm. The first stage is pre-clustering and the second stage is to use the clustering results of the first stage to cluster again. In the pre-clustering stage, the theory of cluster tree growth in BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm [30] is used to process the data points one by one. When processing data points, the clustering tree continually adds and updates a set of split leaf nodes to form many small subclusters [31]. In the second stage of clustering, agglomerative clustering is used to merge and group the preprocessed subclusters. With the Schwarz-Bayesian information criterion and the Akaike information criterion, the optimal number of clusters is determined.

The Euclidean distance function is used to calculate both the degree of dissimilarity between two objects and the degree of closeness and similarity between data individuals, as shown in Equation (3):

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \quad (3)$$

where k = (1,2,3, … … n) represents the internal characteristics of the data individual. In the case of sufficient information, weighting values are assigned to each feature to obtain the weighted Euclidean distance, as shown in Equation (4).

$$d(i,j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \cdots + w_n(x_{in} - x_{jn})^2} \quad (4)$$

For the two-step clustering method, the optimal number of classifications is judged according to the Schwarz Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). In the statistical analysis, the smaller the BIC and AIC values, the better the clustering effect. However, in practice, the BIC change ratio and distance measurement ratio should also be considered. The greater the BIC variation and distance measurement ratio are, the better the clustering effect becomes.

2.4. SDG Model

SDG is a qualitative analysis graph that expresses the interaction between process variables. The directed arc between nodes is helpful to reveal the propagation relationship between variables. The nodes in the model can be physical variables such as pressure and temperature in the system, or operating variables such as valves and controllers. The status values of the nodes are "+", "0", or "−", indicating that its value is greater than the upper threshold, normal state and lower threshold, respectively. If the changing trends of two nodes are the same, that is, the increase of the previous node leads to the increase of the next node, the two nodes are connected by solid arrows. If the trends of two nodes are opposite, that is, the increase of the previous node leads to the decrease of the next node, the two nodes are connected by dotted arrows. A simple SDG model structure is shown in Figure 3. The states of M, N and P are "+", "+" and "·" respectively. The relationship between M and N is represented by a solid arrow, while the relationship between N and P is represented by a dashed arrow, meaning that an increase in M will lead to an increase in N and then a decrease in P [32].

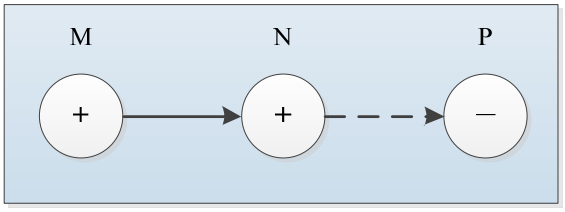

Figure 3. Simple SDG model structure.

The sample of SDG model γ = (G_0, φ) is a function of node state value $\psi : A_0 \to \{+, 0, -\}$. $\psi(n_k)(n_k \in N_0)$ is the symbol of node n_k, as shown in Equations (5)–(7) [24]:

$$\psi(n_k) = 0,\ if\ \left|X_{n_k} - \overline{X}_{n_k}\right| < \varepsilon_{n_k}, \quad (5)$$

$$\psi(n_k) = +,\ if\ X_{n_k} - \overline{X}_{n_k} \geq \varepsilon_{n_k}, \quad (6)$$

$$\psi(n_k) = -,\ if\ X_{n_k} - \overline{X}_{n_k} \leq \varepsilon_{n_k}. \quad (7)$$

where X_{n_k} represents the actual value of the variable corresponding to the node, $\overline{X}_{n_k}$ represents the normal value of the variable corresponding to the node and ε_{n_k} represents the threshold value of the node n_k in the normal state.

3. Industrial Applications

To evaluate the effectiveness of TSCA-SDG, an industrial application is carried out on a catalytic cracking unit for the identification of a reaction temperature anomaly.

3.1. Process Description

The petrochemical catalytic cracking process technology is developed from the thermal cracking process, which can effectively improve the processing depth of crude oil and product quality. It is the core technology for modern refineries to improve heavy distillates and residual oil property. In recent years, due to the shortage of global petroleum resources, the use of petrochemical catalytic cracking technology has become an inevitable trend for petroleum refining companies due to the intensive, energy-saving and environmental protection purposes.

FCC is an important benefit-creating device in the oil refining sector, which can flexibly adjust the product structure. The reaction regeneration system is the core of the catalytic cracking unit, consisting of a reaction part and a regeneration part. The reaction temperature is the main control parameter of the FCC unit, which is an important means to adjust the reaction depth. Increasing the reaction temperature will increase the conversion rate, while the yield and quality of the product will also change. The yield of dry gas and liquid hydrocarbons increase as the reaction temperature rises. However, within different ranges, the range of change is different. The yield of gasoline and diesel has a maximum value with the increase of the reaction temperature. However, after a high value, due to the re-cracking of the products formed by the cracking, a further increase in the reaction temperature will reduce the yield of gasoline and diesel products. Therefore, it is necessary to select an appropriate reaction temperature according to different production schemes.

The process structure of catalytic cracking is complex and its operating environment is harsh, leading to some abnormal working conditions and many unplanned shutdowns. In this paper, the catalytic cracking unit of a petrochemical enterprise is taken as an example for unknown abnormal condition identification. According to the two-year historical data of 347,520 observations in the operation, the abnormal working conditions of the reaction temperature are identified. There are 1700 variables in the whole device, and the data collection cycle is once every three minutes. The collection time for different parts of the process is the same. Abnormal condition identification has far-reaching significance for the long-term stable operation of the device and the improvement of economic benefits.

The flowchart of the catalytic cracking unit is shown in Figure 4. The process consists of three main parts: the reaction regeneration system, the fractionation system and the absorption stabilization system. The reaction regeneration system mainly includes reactor R-101 and regenerator R-102. The fresh oil is mixed with the refining slurry after heat exchange into the lift tube reactor reaction, where FEED1 indicates the vapor extraction steam. The reaction products enter the fractionation system, including feedstock buffer tank D-101, fractionation tower T-101 and diesel vapor extraction tower T-102. Reaction oil and gas enter the fractionation tower from the bottom to the top of the tower. The product at the top of the tower is rich gas and crude gasoline, while the product at the bottom of the tower is oil slurry. OUT1 indicates the side line product light diesel. The absorption and stabilization system mainly consists of absorption tower T-103, reabsorption tower T-106, desorption tower T-104 and stabilization tower T-105. The rich gas and crude gasoline from the fractionation system are separated into liquefied gas OUT2, stabilized gasoline OUT3 and dry gas OUT4 by absorption stabilization, while the rich absorbed oil OUT5 is returned to the fractionation tower.

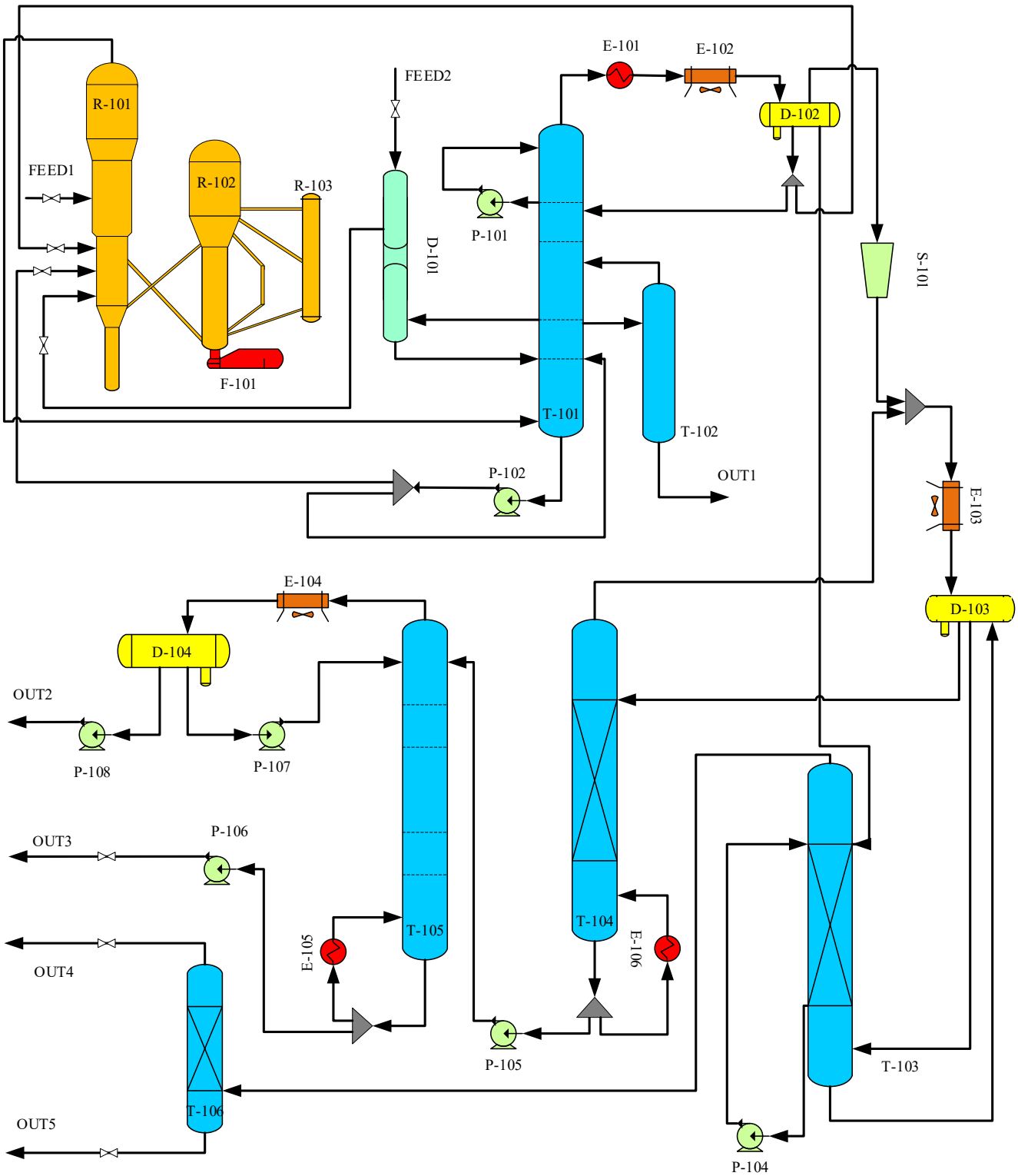

Figure 4. The flowchart of the catalytic cracking unit.

3.2. Data Preprocessing

There are 1700 variables in the whole process of catalytic cracking. However, too many variables will greatly increase the difficulty of unnecessary data analysis. This paper only considered variables related to the reaction temperature. Through communication with field experts, combined with actual work experience, process knowledge and mechanism

analysis, 17 variables were selected. After these 17 variables were standardized by Z Scores standard deviation, the impact of different dimensions was eliminated. The mean value of the converted data is 0, and the standard deviation is 1.

3.3. Feature Extraction and Selection of the Reaction Temperature

The 17 variables are shown in Table 1. Table 1 also lists the Pearson correlation coefficients between these variables and the reaction temperature.

Table 1. Main variables related to the reaction temperature (T1).

Variables	Symbol Description	The Pearson Correlation Coefficient Value with T1
T1	Reaction temperature	1.000
T2	Preheating temperature of the raw materials	0.925
V1	Valve position of the regenerated catalyst slide valve	0.654
F1	Feed quantity	0.944
F2	Slurry oil entering the reactor	0.870
F3	Recycle oil entering the reactor (back under the tower)	0.931
F4	Recycle oil entering the reactor (back on the tower)	0.295
F5	Riser slurry refining line	0.481
F6	Quench water entering the riser	−0.576
F7	Quench oil entering the riser	0.082
F8	Pre-lift dry air volume	0.075
L1	Settler level	0.986
A1	Flue gas oxygen content	0.162
V2	Valve position of the recycle slide valve	0.133
P1	The pressure drop of the regenerated catalyst slide valve	0.874
P2	Dilute phase pressure of the reactor	0.984
P3	Dilute phase pressure of the regenerator	0.986

The variables are clustered by R-type, and the distance between clusters is between-groups linkage distance. The clustering results are shown in Figure 5.

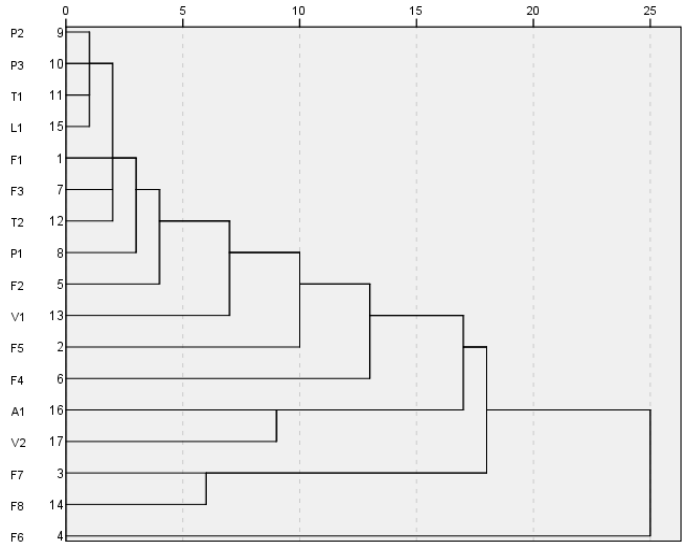

Figure 5. R-type clustering results of the main variables.

Through the results of the correlation analysis in Table 1 and the results of the R-type clustering between variables in Figure 5, the correlation between each variable and the reaction temperature is intuitively reflected. From the ordinate in Figure 5, the order of clustering among variables is given. It can also be seen from the abscissa that these variables are grouped into several classes when given different distances. There are 11 highly correlated variables after feature selection and extraction, as shown in Table 2. After feature selection and extraction, the data are effectively reduced in dimensionality.

Table 2. Variables after feature selection and extraction.

Variables	Symbol Description	The Pearson Correlation Coefficient Value with T1
T1	Reaction temperature	1.000
T2	Preheating temperature of the raw materials	0.925
V1	Valve position of the regenerated catalyst slide valve	0.654
F1	Feed quantity	0.944
F2	Slurry oil entering the reactor	0.870
F3	Recycle oil entering the reactor (back under the tower)	0.931
F6	Quench water entering the riser	−0.576
L1	Settler level	0.986
P1	The pressure drop of the regenerated catalyst slide valve	0.874
P2	Dilute phase pressure of the reactor	0.984
P3	Dilute phase pressure of the regenerator	0.986

3.4. Two-Step Cluster Analysis

The 11 variables obtained were clustered by the two-step clustering method. The BIC automatic clustering results and the AIC automatic clustering results are shown in Tables 3 and 4, respectively. In the rows where the number of clusters in Tables 3 and 4 is 2, the position circled in red box shows that the BIC and AIC values are greatly reduced, and the mutation rate is the smallest and the distance measurement ratio is the largest. The clustering result with good clustering quality is obtained, so the number of clusters is determined to be 2.

Table 3. BIC automatic clustering results.

The Number of Clusters	Schwarz Bayesian Information Criterion (BIC)	BIC Variation	BIC Change Ratio	Distance Measurement Ratio
1	2,649,982.779			
2	575,746.149	−2,074,236.630	1.000	16.391
3	449,464.410	−126,281.739	0.061	1.546
4	367,896.315	−81,568.095	0.039	1.498
5	313,521.815	−54,374.500	0.026	1.533
6	278,154.362	−35,367.453	0.017	1.430
7	253,498.266	−24,656.096	0.012	1.212
8	233,204.125	−20,294.141	0.010	1.016
9	213,232.162	−19,971.963	0.010	1.382
10	198,856.381	−14,375.781	0.007	1.032
11	184,941.682	−13,914.699	0.007	1.142
12	172,795.315	−12,146.367	0.006	1.318
13	163,647.816	−9147.499	0.004	1.293
14	156,638.688	−7009.128	0.003	1.180
15	150,740.056	−5898.632	0.003	1.234

Table 4. AIC automatic clustering results.

The Number of Clusters	Akaike Information Criterion (AIC)	AIC Variation	AIC Change Ratio	Distance Measurement Ratio
1	3,372,404.115			
2	1,181,397.569	−2,191,006.546	1.000	7.252
3	879,331.843	−302,065.726	0.138	1.996
4	728,027.576	−151,304.267	0.069	1.467
5	624,891.975	−103,135.601	0.047	1.287
6	544,778.276	−80,113.698	0.037	1.883
7	502,262.602	−42,515.675	0.019	1.134
8	464,784.110	−37,478.492	0.017	1.076
9	429,958.306	−34,825.804	0.016	1.258
10	402,296.770	−27,661.536	0.013	1.147
11	378,178.655	−24,118.115	0.011	1.079
12	355,836.356	−22,342.299	0.010	1.499
13	340,952.553	−14,883.803	0.007	1.273
14	329,270.612	−11,681.941	0.005	1.096
15	318,620.824	−10,649.788	0.005	1.073

The results of two-step clustering are shown in Table 5. The cluster classes in the two-step clustering results are the first and second classes. The numerical characteristics of class 2 fluctuate in the normal range. Class 2 is, thus, the normal working condition, with 321,856 observations clustered into this class. Compared with the normal value, the numerical characteristics of class 1 have a large fluctuation range. Class 1 is, thus, the abnormal working condition, with 25,664 observations. After cluster analysis, there are data on 25,664 abnormal working conditions.

Table 5. The cluster size of each class of the two-step clustering result.

Clustering Name	Cluster Size	The Number of Observations
Class 1	7.38%	25,664
Class 2	92.62%	321,856

The standardized data of the 11 variables and their two-step clustering results are shown in Figure 6. It can be clearly seen that the two-step clustering has obtained good clustering results. Observation with large fluctuations caused by meter damage, meter calibration, shutdown of the device, etc., are grouped into abnormal working condition class 1.

3.5. Comparison with K-Means Clustering Method

The K-means clustering algorithm uses distance as the evaluation index of similarity. The closer the two data points are, the greater the similarity becomes. Clusters are composed of close data points. The ultimate goal of clustering is to obtain compact and independent clusters.

The K-means clustering algorithm is an iterative solution clustering analysis algorithm. The calculation steps of the clustering algorithm are shown in Figure 7.

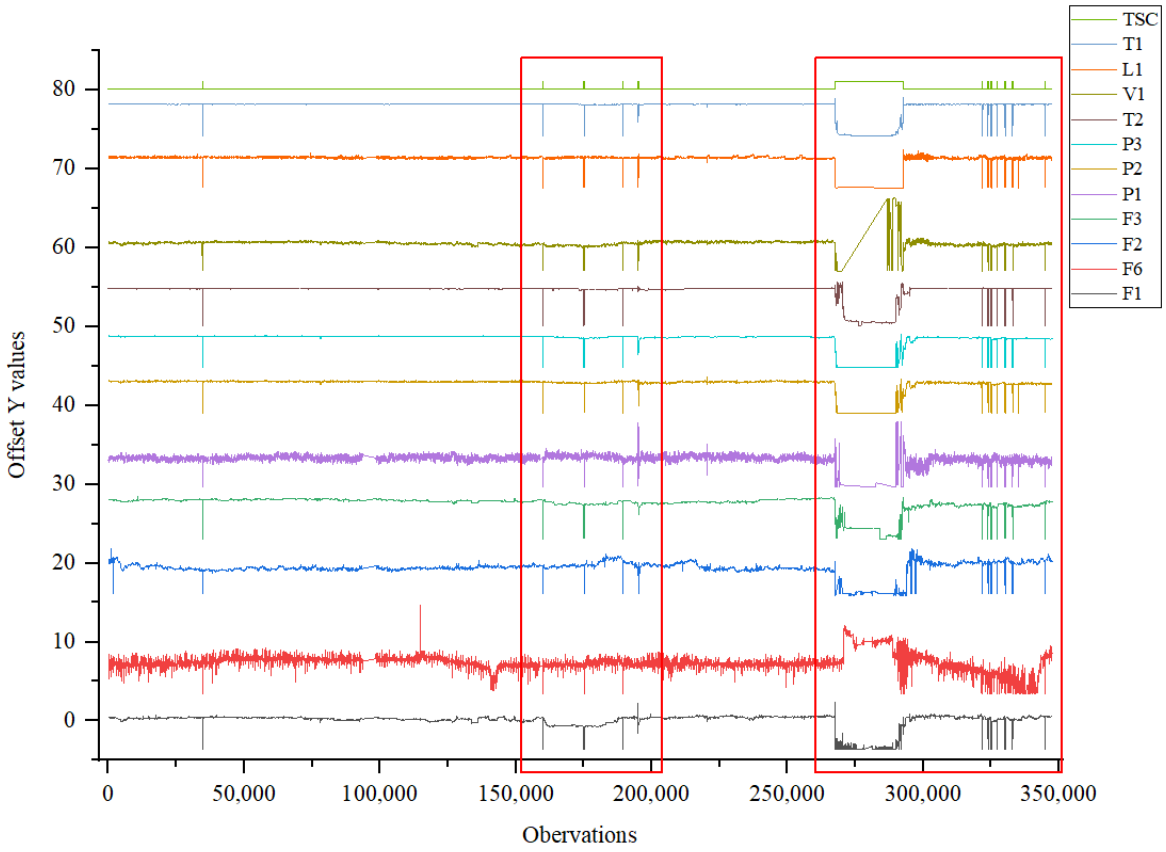

Figure 6. Variables after standardization and two-step clustering results.

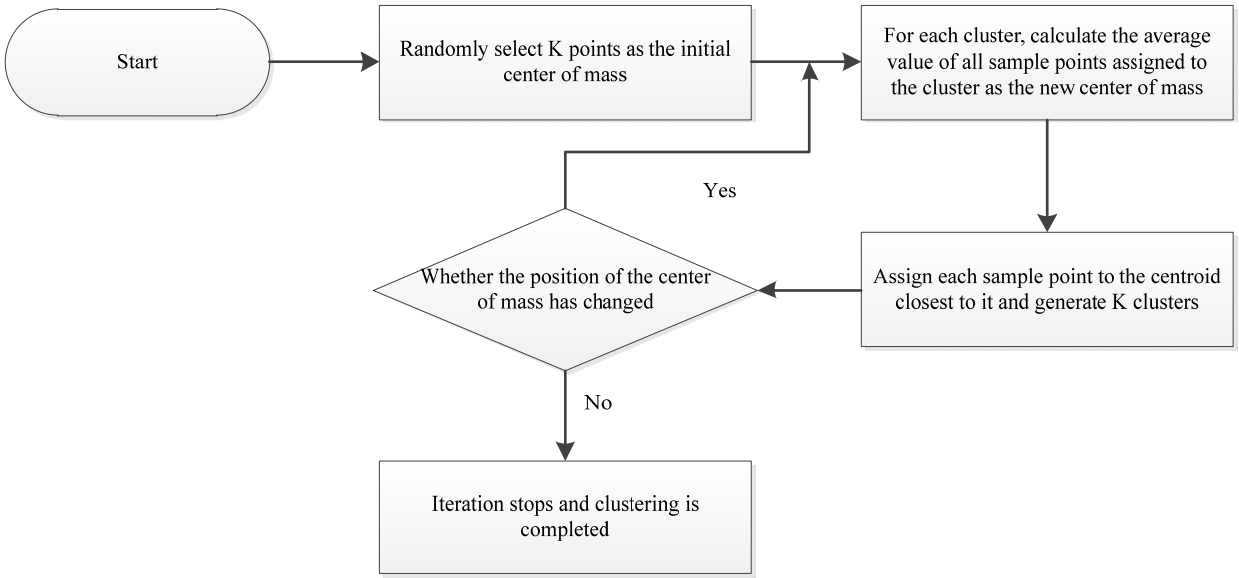

Figure 7. K-means clustering algorithm calculation steps.

In order to facilitate the comparison of clustering results with different clustering methods, the value of K in K-means clustering is set to 2. The K-means clustering results are shown in Table 6.

Table 6. The cluster size of each class of the K-means clustering result.

Clustering Name	Cluster Size	The Number of Observations
Class 1	7.08%	24,607
Class 2	92.92%	322,913

The cluster classes in the K-means clustering results are the first and second classes. Class 2 is the normal working condition, with 322,913 observations clustered into this class. Class 1 is the abnormal working condition, with 24,607 observations. The comparison between the two-step clustering results and the K-means clustering results is shown in Figure 8.

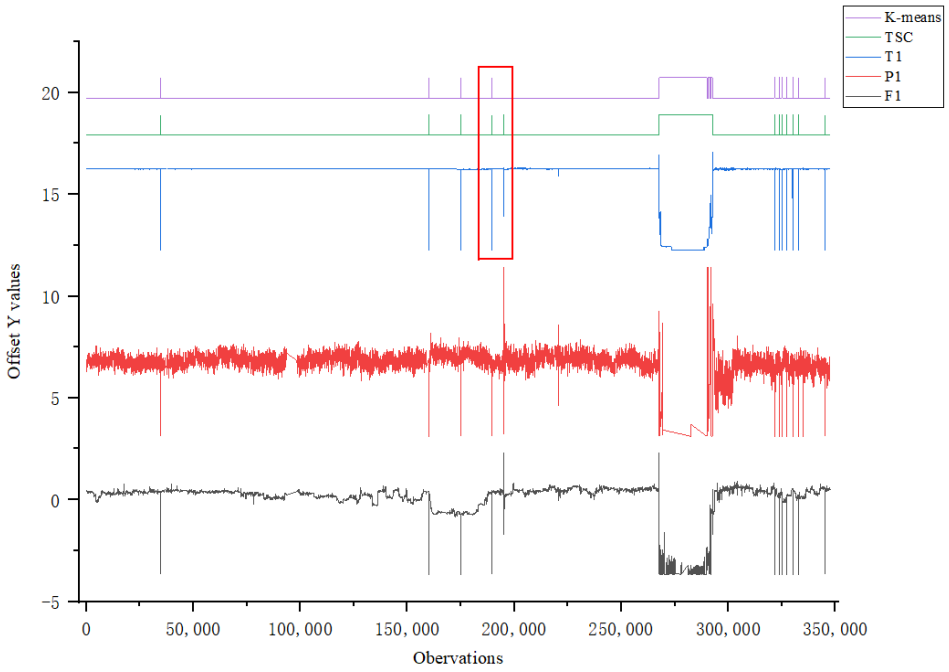

Figure 8. Comparison of the results of the two clustering methods.

In Figure 8, the reaction temperature circled in red fluctuates greatly. The two-step clustering method effectively identifies and classifies them as an abnormal condition, while the K-means clustering method does not identify them effectively. It can, thus, be clearly seen that the two-step clustering method used in this paper is better than the K-means clustering method.

3.6. Establishment of the SDG Model

In actual engineering applications, it is difficult to obtain algebraic equations and differential equations between parameters for large and complex devices and equipment, so the SDG built based on expert experience knowledge is more effective. The use of expert experience alone to establish SDG models has certain limitations, maybe resulting in the inability of system. At the same time, the establishment of SDG model using mathematical analysis alone cannot specifically analyze the relationship between the variables. In this paper, through correlation analysis and R-type clustering analysis, feature selection and extraction are effectively carried out, which has played a key role in reducing the dimensionality of the data variables. By calculating the Pearson correlation coefficient between variables, it effectively and intuitively reflects the correlation between variables. Through the effective combination of expert experience and mathematical analysis, the SDG model of the reaction temperature is well established. The Pearson correlation coefficients among the variables are shown in Table 7.

Table 7. Pearson correlation coefficients among the variables.

	F1	F6	F2	F3	P1	P2	P3	T1	T2	V1	L1
F1	1.000	−0.542	0.825	0.913	0.787	0.938	0.938	0.944	0.899	0.651	0.930
F6	−0.542	1.000	−0.590	−0.455	−0.514	−0.543	−0.544	−0.576	−0.603	−0.387	−0.549
F2	0.825	−0.590	1.000	0.756	0.707	0.850	0.853	0.870	0.800	0.543	0.871
F3	0.913	−0.455	0.756	1.000	0.807	0.938	0.938	0.931	0.901	0.565	0.927
P1	0.787	−0.514	0.707	0.807	1.000	0.879	0.888	0.874	0.830	0.641	0.840
P2	0.938	−0.543	0.850	0.938	0.879	1.000	0.998	0.984	0.909	0.677	0.971
P3	0.938	−0.544	0.853	0.938	0.888	0.998	1.000	0.986	0.909	0.673	0.974
T1	0.944	−0.576	0.870	0.931	0.874	0.984	0.986	1.000	0.925	0.654	0.986
T2	0.899	−0.603	0.800	0.901	0.830	0.909	0.909	0.925	1.000	0.623	0.899
V1	0.651	−0.387	0.543	0.565	0.641	0.677	0.673	0.654	0.623	1.000	0.622
L1	0.930	−0.549	0.871	0.927	0.840	0.971	0.974	0.986	0.899	0.622	1.000

The relationship among the nodes of the reaction temperature related variables is analyzed with mechanism and process data. As shown in Figure 9, the listed influence relationships are combined into a complete SDG model. All nodes are connected by dotted arrows and solid arrows to indicate the negative and positive correlations between the nodes.

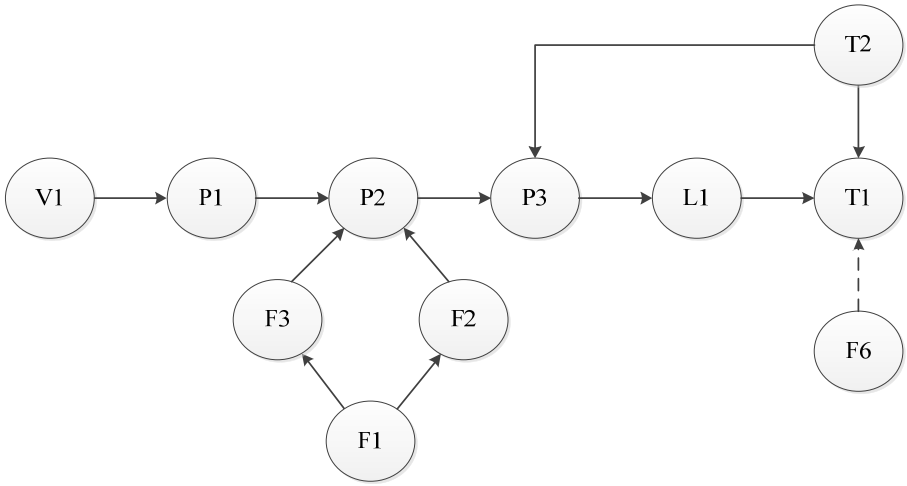

Figure 9. SDG of the reaction temperature.

3.7. Abnormal Identification

In the SDG model, bidirectional inference is used to find the consistent paths and all possible cause nodes. First, the states of all nodes are detected and abnormal nodes that exceed the threshold are found. The transient states of the abnormal nodes are shown in Table 8. Then related nodes are reversely searched from the alarm node T1. Five compatible paths with different cause nodes are obtained as follows:

- T1←L1←P3←P2←P1←V1
- T1←L1←P3←T2
- T1←L1←P3←P2←F2←F1
- T1←L1←P3←P2←F3←F1
- T1←F6

Table 8. States of abnormal nodes.

F1	F6	F2	F3	P1	P2	P3	T1	T2	V1	L1
+	−	+	+	+	+	+	+	+	+	+

The first path is from V1 to T1, and the correlation of each node is positive. The increase of the valve position of the regenerated catalyst slide valve will increase the reaction temperature. The second path from T2 to T1 shows the influence of T2 on T1. The increase in the preheating temperature of the raw materials will also increase the reaction temperature. For the third and fourth paths, the increase of F1 increases F2 and F3, and then T1. In the fifth path, F6 has a negative correlation with T1, of which the increase will lead to the decrease of T1. The SDG model under abnormal conditions is shown in Figure 10. In short, the abnormality of nodes V1, F1, T2 and F6 will cause the fluctuation of reaction temperature T1.

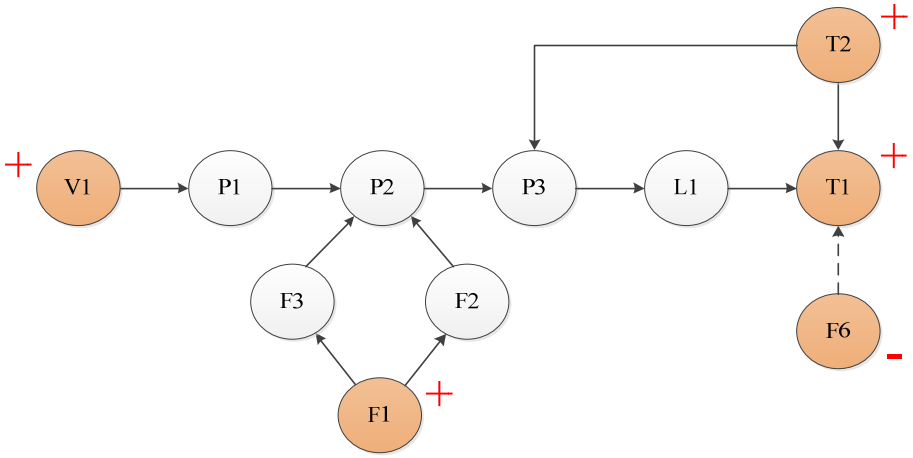

Figure 10. SDG of the reaction temperature.

4. Conclusions

A new TSCA-SDG method is proposed to detect and identify the unknown abnormal working conditions in the catalytic cracking process. Through correlation analysis and R-type clustering analysis, 11 variables are selected, such as feed quantity, preheating temperature of the raw materials and valve position of the regenerated catalyst slide valve. The two-step cluster analysis is performed on 347,520 observations of 11 variables, and the clustering results are obtained as two classes, one for normal working conditions and the other for abnormal operating conditions. The K-means clustering method is used for further verification of the two-step clustering method. SDG model accurately describes the characteristics of abnormal working conditions through the information propagation path between nodes with alarm thresholds. Through the organic combination of cluster analysis with SDG, data dimensionality reduction and feature selection and extraction are effectively carried out. Then, abnormal working conditions are quickly identified. From the perspective of mechanism analysis, the identification of unknown abnormal working conditions in catalytic cracking is better and more accurate than experience only. At present, there is much research on the identification of known working conditions, while there is little research on the identification of unknown working conditions, although this is urgently needed because there are a lot of abnormal working conditions in industrial production. The TSCA-SDG method proposed in this paper solves this problem meaningfully. The quality of the clustering algorithm will limit the identification of abnormal conditions, so its further development will promote more in-depth research.

Author Contributions: Methodology, J.H.; formal analysis, J.H., Y.L., Z.C., Z.L. and C.L.; writing—original draft preparation, J.H. and J.Q.; wring—review and editing, J.H., J.Q., W.T. and Z.C.; validation, J.H., W.T., Y.L. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 22178189.

Data Availability Statement: Not applicable.

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

References

1. Clough, M.; Pope, J.C.; Lin, L. Nanoporous materials forge a pathforward to enable sustainable growth: Technology advancements in fluid catalytic cracking. *Microporous Mesoporous Mater.* **2017**, *254*, 45–58. [CrossRef]
2. Souza, N.L.A.; Tkach, I.; Morgado, E.; Krambrock, K. Vanadium poisoning of FCC catalysts: A quantitative analysis of impregnated and real equilibrium catalysts. *Appl. Catal. A Gen.* **2018**, *560*, 206–214. [CrossRef]
3. Salvado, F.C.; Teixeira-Dias, F.; Walley, S.M.; Lea, L.J.; Cardoso, J.B. A review on the strain rate dependency of the dynamic viscoplastic response of FCC metals. *Prog. Mater. Sci.* **2017**, *88*, 186–231. [CrossRef]
4. Yang, F.; Zhou, M.; Jin, J.M.; Cao, M. Research Progress on Application of Intelligent Optimization Algorithm. *Acta Pet. Sin. (Pet. Process. Sect.)* **2020**, *36*, 878–888.
5. Venkatasubramanian, V.; Rengaswamy, R.; Yin, K.; Kavuri, S.N. A review of process fault detection and diagnosis Part I: Quantitative model-based methods. *Comput. Chem. Eng.* **2002**, *27*, 293–311. [CrossRef]
6. Cong, X.Y.; Fanti, M.P.; Mangini, A.M.; Li, Z.W. Decentralized fault diagnosis by Petri nets and integer linear programming. *Ifac Pap.* **2017**, *50*, 13624–13629. [CrossRef]
7. Xu, Y.; Deng, X.G.; Zhong, N. A fault diagnosis method for multimode processes based on ICA mixture models. *CIESC J.* **2016**, *67*, 3793–3803.
8. Gururajapathy, S.S.; Mokhlis, H.; Illias, H.A. Fault location and detection techniques in power distribution systems with distributed generation: A review. *Renew. Sustain. Energy Rev.* **2017**, *74*, 949–958. [CrossRef]
9. Ge, Z.Q.; Song, Z.H.; Gao, F.R. Review of Recent Research on Data-Based Process Monitoring. *Ind. Eng. Chem. Res.* **2013**, *52*, 3543–3562. [CrossRef]
10. He, C.; Ge, D.C.; Yang, M.H.; Yong, N. A data-driven adaptive fault diagnosis methodology for nuclear power systems based on NSGAII-CNN. *Ann. Nucl. Energy* **2021**, *159*, 108326. [CrossRef]
11. Li, X.; Liu, J.Y.; Liu, B.; Zhang, Q.; Li, K.N.; Dong, Z.X.; Mou, L.J. Impacts of data uncertainty on the performance of data-driven-based building fault diagnosis. *J. Build. Eng.* **2021**, *43*, 103153. [CrossRef]
12. Jiang, L.L.; Deng, Z.W.; Tang, X.L.; Hu, L. Data-driven fault diagnosis and thermal runaway warning for battery packs using real-world vehicle data. *Energy* **2021**, *234*, 121266. [CrossRef]
13. Wang, Z.J.; Zhao, W.L.; Du, W.H.; Li, N.P.; Wang, J.Y. Data-driven fault diagnosis method based on the conversion of erosion operation signals into images and convolutional neural network. *Process Saf. Environ. Prot.* **2021**, *149*, 591–601. [CrossRef]
14. Yao, Y.M.; Luo, W.J.; Dai, Y.Y. Research progress of data-driven methods in fault diagnosis of chemical process. *Chem. Ind. Eng. Prog.* **2021**, *40*, 1755–1764.
15. Venkatasubramanian, V.; Rengaswamy, R.; Ka, S.N.; Kavuri, S.N.; Ka, S.N. A Review of Process Fault Detection and Diagnosis Part II : Qualitative Models and Search Strategies. *Comput. Chem. Eng* **2003**, *27*, 313–326. [CrossRef]
16. Maurya, M.R.; Rengaswamy, R.; Venkatasubramanian, V. Fault Diagnosis by Qualitative Trend Analysis of the Principal Components. *Chem. Eng. Res. Des* **2005**, *83*, 1122–1132. [CrossRef]
17. Alauddin, M.; Khan, F.; Imtiaz, S.; Ahmed, S. A Bibliometric Review and Analysis of Data-Driven Fault Detection and Diagnosis Methods for Process Systems. *Ind. Eng. Chem. Res.* **2018**, *57*, 10719–10735. [CrossRef]
18. Tan, J.P.; Yang, Z.J.; Cheng, Y.Q.; Ye, J.L.; Wang, B.; Dai, Q.Y. SRAGL-AWCL: A two-step multi-view clustering via sparse representation and adaptive weighted cooperative learning. *Pattern Recognit.* **2021**, *117*, 107987. [CrossRef]
19. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
20. Wang, L.M.; Wang, H.H.; Han, X.M.; Zhou, W. A novel adaptive density-based spatial clustering of application with noise based on bird swarm optimization algorithm. *Comput. Commun.* **2021**, *174*, 205–214. [CrossRef]
21. Zeng, H.; Cheung, Y.M. A new feature selection method for Gaussian mixture clustering. *Pattern Recognit.* **2009**, *42*, 243–250. [CrossRef]
22. Murtagh, F.; Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.* **2014**, *31*, 274–295. [CrossRef]
23. Yang, X.J.; ZHOU, Y.Y.; LIU, M.D. The Fusion Recognition Using Fingerprints and Handwritten Signature Based on the Bayesian Algorithm. *Int. Conf. Artif. Intell. Softw. Eng.* **2013**, 147–151. [CrossRef]
24. Yang, F.; Xiao, D.Y. Review of SDG modeling and its application. *Control Theory Appl.* **2005**, *5*, 93–100.
25. Gao, D.; Xu, X. Signed Directed Graph and Qualitative Trend Based Model Semiquantitative Validation. *Control Eng. China* **2019**, *26*, 515–520.
26. Wu, G.H.; Yuan, D.P.; Yin, J.Y. A Framework for Monitoring and Fault Diagnosis in Nuclear Power Plants Based on Signed Directed Graph Methods. *Front. Energy Res.* **2021**, *9*, 641545.
27. Guo, X.; Zhao, L.; Homayoun, H.; Dinakarrao, S.M.P. Deep graph transformation for attributed, directed, and signed networks. *Knowl. Inf. Syst.* **2021**, *63*, 1305–1337. [CrossRef]
28. Du, Q.; Jia, L.Y. *SPSS Statistical Analysis from Entry to Proficiency*; Posts&Telecom Press: Beiijng, China, 2009; pp. 257–258.
29. Xue, W. *Statistical Analysis and SPSS Application*; China Renmin University Press: Beiijng, China, 2017; pp. 269–272.

30. Madan, S.; Dana, K.J. Modified balanced iterative reducing and clustering using hierarchies (m-BIRCH) for visual clustering. *Pattern Anal. Appl.* **2016**, *19*, 1023–1040. [CrossRef]
31. Zhang, Y.M.; Huang, Y.S. Analysis of Passenger Characteristics of Regular Bus Passengers Based on Two Step Cluster Algorithm-A Case Study of Xiamen. *China Transp. Rev.* **2020**, *42*, 120–126.
32. Tian, W.D.; Zhang, S.F.; Cui, Z.; Liu, Z.J.; Wang, S.C.; Zhao, Y.; Zou, H. A Fault Identification Method in Distillation Process Based on Dynamic Mechanism Analysis and Signed Directed Graph. *Processes* **2021**, *9*, 229. [CrossRef]