Semi-Supervised Fuzzy C-Means Clustering Optimized by Simulated Annealing and Genetic Algorithm for Fault Diagnosis of Bearings

JIANBIN XIONG, (Member, IEEE), XI LIU, (Member, IEEE), HONGBIN ZHU, HAIIYING LI, and QINGHUA ZHANG, (Member, IEEE)

School of Automation, Guangdong Polytechnic Normal University, Guangzhou 510665, China
School of Computer, Guangdong University of Petrochemical Technology, Maoming 525000, China
Department of Automatic Control, School of Automation, Guangdong University of Technology, Guangzhou 510006, China
Department of Economics and Trade, Guangdong Maoming Agriculture and Forestry Technical College, Maoming 529024, China
Guangdong Province Petrochemical Equipment Fault Diagnosis (PEFD) Key Laboratory, Guangdong University of Petrochemical Technology, Maoming 525000, China

Corresponding author: Xi Liu (liuxi7790@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 62073090 and 61473331, in part by the Natural Science Foundation of Guangdong Province of China under Grant 2019A1515010700, in part by the Key (Natural) Project of Guangdong Provincial under Grant 2019KZDZX020 and Grant 2019KZDZX1004, in part by the Introduction of Talents Project of Guangdong Polytechnic Normal University of China under Grant 991512203 and Grant 991560236, and in part by the Guangzhou Key Laboratory Project of Intelligent Building Equipment Information Integration and Control under Grant 202002010003, in part by the Key Project of Ordinary University of Guangdong Province under Grant 2019KZDZX020, 2020KZDZX2014, in part by the Guangzhou People’s Livelihood Science and Technology Project under Grant 201903010059, and in part by the Guangzhou Yuelu District Science and Technology Plan Major Project under Grant 2019-GX-010, and in part Guangdong University Students Science and Technology Innovation Cultivation Special Fund Subsidy Project under Grant pdjh20200333.

ABSTRACT As a popular clustering algorithms, fuzzy c-means (FCM) algorithm has been used in various fields, including fault diagnosis, machine learning. To overcome the sensitivity to outliers problem and the local minimum problem of the fuzzy c-means new algorithm is proposed based on the simulated annealing (SA) algorithm and the genetic algorithm (GA). The combined algorithm utilizes the simulated annealing algorithm due to its local search abilities. Thereby, problems associated with the genetic algorithm, such as its tendency to prematurely select optimal values, can be overcome, and genetic algorithm can be applied in fuzzy clustering analysis. Moreover, the new algorithm can solve other problems associated with the fuzzy clustering algorithm, which include initial clustering center value sensitivity and convergence to a local minimum. Furthermore, the simulation results can be used as classification criteria for identifying several types of bearing faults. Compare with the dimensionless indexes, it shows that the mutual dimensionless indexes are more suitable for clustering algorithms. Finally, the experimental results show that the method adopted in this paper can improve the accuracy of clustering and accurately classify the bearing faults of rotating machinery.

INDEX TERMS Rotating machinery, mutual dimensionless indexes, fuzzy c-means clustering algorithm, genetic algorithm, simulated annealing algorithm.

I. INTRODUCTION

Rotating machinery is the most popular type of equipment used in mechanical engineering industrial applications. Rolling bearings are one of the most important components of rotating machinery. However, the rapid developments in science and technology have increased the complexity of rotating machinery structures, increasing the probability of rolling bearing failure [1]. Therefore, rotating machinery diagnosis for predicting rolling bearing failures is of particular significance [2], [3]. A common fault diagnosis method is to analyze vibration signals that contain mechanical fault information [4], [5]. This method comprises two important steps: signal feature extraction and fault status identification. Due to the fact that mechanical failures occur gradually, there is often uncertainty in relation to which vibration signal characteristics should be extracted, and a fault state may not be identified from the eigenvalues extracted from the vibration signals. Thus, an effective way to identify failure modes is to use fuzzy c-means (FCM) clustering based on the fuzzy theory [6], [7].
Significant research for rotating machinery fault diagnosis has been previously conducted to solve the problem of complex rotating machinery failure. To date, a number of efficient fault diagnosis methods have been derived. To improve the dimensionless index classification of petrochemical rotating machinery equipment, a double sample data fusion method is proposed by Xiong et al. Based on a combination of rules and raw data collection, each dimensionless index could be used as the evidence of the system, and then, the Kolmogorov-Smirnov (K-S) test could be used to detect the exact type of failure. Their experiments demonstrated successful fault type identification using dimensionless indexes with coincidences or evidence conflicts. Compared to the k-nearest neighbors (KNN) algorithm, this method provided better fault recognition and improved the fault detection accuracy by 9.45% [8]. Under actual working conditions, serious overlap problems can occur when the dimensionless parameter range is defined by vibration monitoring calculations to simulate compound faults. On this basis, Sun et al. proposed a novel online method based on dimensionless immune detector and evidential reasoning (ER) to diagnose rotating machinery. Their method was able to effectively achieve real-time fault diagnosis with great potential for practical engineering applications [9].

FCM is a popular clustering method using the concept of geometric closeness of the determined data point to classify data [10]. At present, clustering algorithms are widely used in different domains, such as mechanical fault diagnosis, medical data processing, and image processing [11]–[13]. Wang et al. applied the fuzzy c-means clustering algorithm in mechanical fault diagnosis. The center of the data eigenvalues of rolling bearings was used to distinguish the fault categories and good results were obtained. However, when processing complex data, the above algorithm may not be sufficient when using a certain point as a clustering center for a certain type of data [14].

When the data eigenvalues are difficult to be separated, the number of different types of data differs, or the distance between different data types is small, it is convenient to shift the data center point to a category with more data, which will reduce the accuracy of clustering. In such cases, it is important to select the initial clustering center point. According to the above analysis, Wu et al. proposed an improved algorithm, which takes into account two factors, the distance and the local density, and it was able to get good results [15].

Zhang et al. gave a FCM algorithm based on a genetic algorithm (GA). It was shown that when the number of samples and categories is large, the proposed algorithm can lead to faster speed and more accurate results than the FCM algorithm. Bai et al. proposed a fuzzy clustering algorithm based on simulated annealing (SA) and genetic algorithm. The above algorithm could overcome the system sensitivity to data sets and initial clustering centers, and avoid falling into local minima [16].

Among the aforementioned algorithms, the optimization algorithms proposed by Zhang et al. and Bai et al. let the sum of the distances between the clustering center and the data points be the fitness function in genetic algorithm. It is generally believed that the smaller the distance, the higher the fitness. However, this method may not be appropriate for processing some complex data types, especially when the data does not satisfy normal distribution. In addition, this algorithm is sensitive to outliers.

In addition, semi-supervised clustering, a new learning method, is proposed in recent years, which combines semi-supervised learning and cluster analysis [17], [18]. The exiting semi-supervised clustering algorithm can be divided into three class, such as the method based on constraint, the method based on distance, and the method based on constraint and distance [19].

To overcome the shortcoming of the method proposed by Zhang et al, in this paper, the distance and clustering accuracy are taken as the fitness function of genetic simulated annealing algorithm. The following contributions are presented:

1) A mutual dimensionless method is used to process data. The mutual dimensionless index can reduce the distance between the internal structure of each dimensionless index and the coincidence of the same dimensionless index of each fault. The experimental results show that, compare with dimensionless method, the accuracy of bearing fault diagnosis can be increased by 9.22% at most by mutual dimensionless processing.

2) Since traditional FCM is very sensitive to outlier, the clustering accuracy is taken as the objective function, which is optimized by the genetic simulation annealing algorithm.

3) Several popular intelligent fault diagnosis methods, including the GA, SA, and FCM clustering algorithm, are merged to form an integrated diagnosis method for rotating machinery.

The paper is organized as follows: Section II will introduce our related work on data processing and fault diagnosis. Section III provides the FCM clustering theory, the GA and SA algorithm principles, and the theory of the fusion algorithm for single intelligent fault diagnosis. Section IV presents the experimental procedure, results, and analysis on the hybrid fault diagnosis method based on the SA and GA algorithms. Section V provides a general discussion, while a summary of the study and future directions are presented in Section VI.

II. RELATED WORK

This section describes the related work on data processing of rotating unit and our previous for fault diagnosis. Our related works show the feasibility of our algorithm to some extent.

A. RELATED WORK ON DATA PROCESSING

The characteristic signals of the bearing are different, when the it is in different operational state. The original vibration signal can be analyzed and processed, which can better meet the needs of state diagnosis. For example, time domain waveform signal is a kind of original vibration signal without any processing. Its waveform characteristics are different with different operation status, as shown in the Fig. 1 to 5,
However, it is difficult for us to identify the type of the operation status. Therefore, it is feasible and significant to extract fault feature from time domain signals and dimensional index often used in time domain analysis.

Xiong et al. proposed a method based on the static discount factor, combining the KNN classification algorithm with the dimensionless index for information fusion fault diagnosis, and experimental results show the method can reduce the impact of unreliable factors on the fusion effect. But there are difficulties with distinguishing evidence from complete fusion or large conflicts resulting in uncertainty in the problem diagnosis [20].

Then, Xiong et al. exchanged numerator and denominator in dimensionless formula, and named it the mutual dimensionless index. It is proved in reference [21] that the mutual dimensionless index is sensitive enough to the fault and not affected by the working conditions of the machine, which is more suitable for the fault diagnosis of rotating machinery. Accordingly, the mutual dimensionless method is used to process the data in this paper.

### B. RELATED WORK ON FAULT DIAGNOSIS

Xiong et al. proposed SVM and correlation coefficient algorithm to diagnose mechanical bearing, according to the separability and correlation of fault data eigenvalues [8], [21]. However, the method of SVM did not solve the problem of parameters selection, and the method of correlation coefficient has the problem of low diagnosis rate.

Wang et al. used the FCM algorithm to diagnosis mechanical status [14]. The center of data eigenvalues can be obtained by FCM. However, the method does not overcome the local minimum problem of FCM. Therefore, this paper improves the FCM algorithm by using SA and GA.

### III. RELATED THEORY

For readability and ease of later development, this section will describes three algorithms used in the paper, including fuzzy c-means clustering algorithm, genetic algorithm and simulated annealing algorithm.

#### A. FUZZY C-MEANS CLUSTERING ALGORITHM

The FCM clustering algorithm is one of most popular clustering algorithms, which is based on the objective function [22]. Known data points are grouped into various categories, and the cluster centers and degree of membership in each category are calculated [23]. The objective function is optimized by calculating the membership of the given data points to all cluster centers. However, the algorithm is very sensitive to the parameter initial state values, which means that it can easily
fall into a local minimum point [24]. Using a combination of the SA and GA algorithms, the function can quickly converge to the global optimal solution, where the objective-function value of the non-similarity index is stabilized at its minimum.

Hypothesis 1: There is a dataset \( x = (x_1, x_2, \ldots, x_n) \) which is grouped into \( C \) categories, \((S_1, S_2, \ldots, S_C)\), where \( C \in [2, n] \). Each sample has \( C \) membership degrees and \((N_1, N_2, \ldots, N_C)\) are the cluster centers of \((S_1, S_2, \ldots, S_C)\).

Theorem 1: The objective function \( J_m \) can be expressed as follows:

\[
J_m(M, N) = \sum_{i=1}^{n} \sum_{j=1}^{c} A_{ij} b d_{ij}^2
\]  

(1)

where \( A_{ij} \) is the membership of the sample \( x_i \) relative to category \( S_j \).

Lemma 1: The Euclidean distance can be expressed as follows:

\[
d_{ij} = d(x_i - N_j) = \frac{1}{2} \left( \sum_{k=1}^{m} (x_{ik} - N_{jk})^2 \right)
\]  

(2)

where \( d_{ij} \) is the Euclidean distance, defined as the distance between the \( k^{th} \) sample \( x_i \) and the center \( N_j \) of the \( j^{th} \) class.

Lemma 2: The purpose of fuzzy clustering is to find an optimal membership function \( M \) to minimize the value of the objective function \( J_m \). The total of the membership degrees of all samples is:

\[
\sum_{k=1}^{c} A_k(x_i) = 1, \quad i = 1, 2 \ldots, n.
\]  

(3)

Then, the membership degree of the sample point \( x_i \) relative to class \( S_j \) can be calculated as:

\[
A_{ij} = \left( \sum_{k=1}^{c} \frac{d_{ij}}{(7)} \right)^{-1}, \quad i = 1, 2 \ldots, n.
\]  

(4)

The clustering center \( (N_i) \) can be calculated as:

\[
N_{ij} = \frac{\sum_{j=1}^{c} (A_{ij})^b x_{ij}}{\sum_{j=1}^{c} (A_{ij})^b}.
\]  

(5)

Equation (5) used to calculate \( c \) cluster centers. Equations (4) and (5) are used to adjust the clustering center and membership degree repeatedly. Therefore, the clustering centers of all samples types and the membership degrees of each sample can be obtained theoretically, and thus, the partition of the fuzzy clustering algorithm is completed. Despite that the FCM has a high retrieval speed, it is a local retrieval algorithm which is sensitive to the initial clustering center. The algorithm will fall into a local minimum, if the initial value is not properly selected [10].

Fuzzy c-means clustering algorithm, using fuzzy theory, obtains the membership degree of each sample point to all class centers by optimizing the objective function, and weights the membership degree, so as to determine the class of sample points to classify sample data. This method gets the minimum value of \( J_m \) function, but the result may be the local minima or saddle point of the function. Besides, FCM is easily affected by outliers. Therefore, fuzzy c-means clustering algorithm is not suitable for the bearing fault of large-scale petrochemical units with noise and sample imbalance. To solve the above problems, this paper proposes the application of semi supervised fault diagnosis method based on simulated annealing and genetic algorithm optimization in bearing fault diagnosis. See parts B and C for details.

B. GENETIC ALGORITHM

The GA, proposed by John Henry Holland in the 1870s, is a computational model that mimics natural evolutionary systems, and uses the biological evolution theory and the stochastic exchange theory according to Darwin’s survival of the fittest [25]. The basic concept is to randomly generate an initial population, whose individual genotype is a tree structure. Then, according to the survival of the fittest principle, population duplication, crossover, and mutation are performed using iterative optimization. Based on the size of the fitness value of each individual, the best individuals are selected, forming a new population. The purpose of the iterative process is to make the offspring population adapt better to the environment and, once an iteration is terminated, to decode the optimal individuals as the optimal solution, which is usually used for parameter selection [26], [27]. However, in the early stages of the iterative process, GA can easily cause the whole population to consist of super individual offspring, leading to premature optimization. Since in the later stages of the GA, the individual population fitness will be similar, the predominance of “super individuals” is more apparent in the offspring [28]. A detailed flowchart is shown in Fig. 6, where GEN is the population size of the counter and MAXGEN is the maximum number of evolutions.

1) Coding method: In this paper, the number of parameters to be optimized are \( C \) initial clustering centers, in which each dyeing is composed of \( C \) clustering centers. Assume that each variable has M-dimensional and is encoded by \( K \) binary codes. The length of chromosome is \( C \times M \times K \).

2) Fitness function: It is used to evaluate the degree of adaptation of each code string. The search type in genetic algorithms is guided only by the fitness function \( \text{Fit} \) [29]. The Fit in the paper is the accuracy after clustering and \( J_m \) in Equation (1). The bigger the accuracy is, the smaller the reciprocal is, and the bigger the fitness is. The definition of fitness in this article is

\[
\text{Accuracy} = \frac{\sum_{i=1}^{c} h_i}{\sum_{i=1}^{c} H_i}
\]  

(6)

\[
\text{Fit} = J_m + \frac{1}{\text{Accuracy}}.
\]  

(7)

where \( h_i \) denotes the number of data belonging to class \( i \) after clustering, and \( H_i \) denotes the total number of data belonging to class \( i \).

3) Selection operator: Its function is to copy parent chromosomes of high fitness to the following generation.
Assuming a population size of $U$, the probability $P$ that the individual $t$ can be chosen is

$$P = \frac{\text{Fit}(t)}{U}. \quad (8)$$

4) Variation operator: It is an auxiliary method for generating new individuals, ensuring genotype diversity in the population, while preventing search stagnation [29].

Genetic algorithm takes the fitness function as the information guiding the search type. It only uses fitness function value to measure the excellence of an individual, and does not involve the process of derivation and differentiation of objective function value, which makes genetic algorithm show a high degree of superiority, because in reality, many objective functions are difficult to derive, or even do not have derivatives. However, with the extensive research, the disadvantages of genetic algorithm have come to light, such as:

1) The genetic algorithm is prone to nonstandard and inaccurate problems in coding.

2) As a single genetic algorithm coding can not fully express the constraints of the optimization problem, it is necessary to consider the threshold value for the infeasible solution, which increases the workload and solution time.

3) The efficiency of genetic algorithm is usually lower than other traditional optimization algorithms.

4) Genetic algorithm is prone to premature convergence.

C. SIMULATED ANNEALING ALGORITHM

To overcome the shortcomings of genetic algorithm, simulated annealing algorithm is used in this paper. The SA algorithm was first proposed by Metropolis et al. in 1953 [30] and was successfully introduced into the optimization problem-solving process by Kirkpatrick et al. of the IBM Thomas W. Watson Research Center in 1983 [31]. The basic principle of SA, which is used to simulate the optimal solution process, is based on the solid annealing process in physics, where a solid is first heated to a certain temperature, so as to melt it, and then it is slowly cooled into a solid with regular microstructure. As the temperature of the solid increases, the particles in the solid body accelerate and move continuously in a non-uniform manner, and as the solid cools down, they decelerate. When the temperature reaches the ambient temperature, the particles are in the lowest energy state and in a state of thermal equilibrium [32]. In the SA algorithm, the solution is obtained by selecting a non-local optimal solution with a certain probability that falls within the scope of the initial solution. The loop iterates and as the simulated temperature decreases, the algorithm activity declines, approaching eventually the global optimum.

According to the Metropolis criterion, the probability $P$ of a particle becoming stable when the temperature of the particle is $T$ can be defined by:

$$P(\Delta E) = e^{-\frac{\Delta E}{k_b T}} \quad (9)$$

where $E$ is the internal energy of the particle at temperature $T$, $\Delta E$ is the change in internal energy, and $k_b$ is the Boltzmann constant. At temperature $T$, the particle satisfies the Boltzmann probability distribution [33]–[35], which for a stable molecule in the $x$ state is derived by:

$$P(\tilde{E} = E(x)) = \frac{1}{Z(T)} e^{-\frac{E(x)}{k_b T}} \quad (10)$$

where $\tilde{E}$ is a random molecular energy variable, $E(x)$ denotes the energy of the molecule in the $x$ state, and $Z(T)$ is the normalized factor of the probability distribution, which can be defined as:

$$Z(T) = \sum_{x \in D} e^{-\frac{E(x)}{k_b T}} \quad (11)$$

Steps to implement the SA algorithm are shown as follows.  
1) The initial solution $M_0$ is arbitrarily generated. Let $M(0) = M_0$, set the initial temperature as $T_0$, and let $T = T_k$, where $k = 0$.

2) Use the Metropolis criterion to assess $M_k$ and $T$, then return $M_i = M$, with $M_k$ as the current solution.

3) Cool down to the current temperature $T$ and let $T = T_k + 1$. If $T_k + 1 < T_k$, add 1 to $k$ ($k = k + 1$).

4) Check if the annealing algorithm satisfies the termination condition. If yes, continue to step 5), otherwise, return to step 2).

5) Use $M_k$ as the current optimal solution, output the optimal value, and end the algorithm.

The advantages of SA are the following:

1) SA can deal with the objective function with any degree of nonlinearity, discontinuity and randomness, and the objective function can have any boundary conditions and constraints.
2) Compared with other linear optimization methods, SA and is easy to implement with less programming work.

3) It can guarantee to find the global optimal solution statistically.

The disadvantages of SA are the following:

1) It takes a lot of time to find the optimal solution, especially when using “standard” Boltzmann sampling technology (i.e. standard receiving function).

2) Compare with other algorithms, It needs more difficult parameter adjustment to solve a specific problem.

3) Cooling too fast will leads to simulated annealing to simulated quenching (SQ), which is not statistically guaranteed to find the optimal solution.

IV. CLUSTERING ALGORITHM BASED ON SIMULATED ANNEALING AND GENETIC ALGORITHM

The three algorithms described above, are associated with some disadvantages. In the proposed GA-SA-FCM algorithm the following improvements have been made:

1) In order to solve the poor searching ability problem of the GA, this study adopts the strong local searching ability of the SA algorithm. The SA algorithm is integrated into the solving process of the GA, in order to avoid the local optimal searching process of the GA.

2) The combination of the GA and the SA algorithm can significantly improve the global search ability of the whole algorithm.

3) This paper combines the SA with GA, and then, applied to the FCM. By using the strong local search ability of the SA and the strong global search ability of the GA, the clustering problem can be solved effectively and quickly [36].

Firstly, set the parameters for three algorithms. Secondly, cluster the training data, through which the distance value can be got. Besides, we can also get the accuracy after clustering by comparing the clustering category with the actual category. Then the distance and the reciprocal of the accuracy value were taken as the fitness function of the genetic algorithm. Here, the combination algorithm of SA and GA were used to adjusts the clustering center of clustering algorithm until the algorithm converges. Finally, a superior clustering center can be obtained. The flow chart of the proposed FCM based on the GA and SA algorithms is shown in Fig. 7.

A. STEPS TO IMPLEMENT THE FUZZY C-MEANS CLUSTERING BASED ON SIMULATED ANNEALING AND GENETIC ALGORITHM

begin

setup: 1) Initialize the FCM clustering parameters: the power index is 3, the maximum number of iterations is 20, the objective function termination tolerance is $1 \times 10^6$.

2) Initialize the SA algorithm parameters: the cooling coefficient $q$ is 0.8, the initial temperature $T_0$ is 100, and the termination temperature $T_{end}$ is 1.

3) Initialize the genetic algorithm parameters: the number of individual sizepop is 10, the maximum genetic algebra
TABLE 1. Operation status and number of the normal and fault bearing categories.

| Label       | Label 1            | Label 2        | Label 3          | Label 4                  |
|-------------|--------------------|----------------|------------------|--------------------------|
| Machine operation number | Normal bearing | Outer-ring wear | Inner-ring wear | Bearing with one ball less |
| 98          | 98                 | 98             | 98               | 98                       |

MAXGEN is 10, the variable dimension, the variable binary digit PRECI is 10, the generation gap GGAP is 0.9, the crossover probability pc is 0.7, and the mutation probability pm is 0.01.

4) Create the initial population Chrom, calling the objective function ObjFun, to calculate the objective function value of the initial population individuals ObjV.

The detail steps to implement the GA-SA-FCM Algorithm is shown in Algorithm 1.

Algorithm 1 The Detail Steps to Implement the GA-SA-FCM Algorithm

```plaintext
do
    while T > T_{end} do
        Initialize generation counte GEN = 0
        while GEN < MAXGEN do
            Genetic algorithm selection, reorganization, mutation operation, and calculate the target function objVSele offspring, reinsert the establishment of a new population newChrom, new target function value newObjV
            for i : sizepop do
                if current objective function value ObjV(i) > new objective function then
                    Calculate newObjV(i)
                    Replace the old individual
                else
                    Randomly generated probability P
                    if P ≤ \exp(\frac{\text{ newObjV}(i) - \text{ ObjV}(i)}{T}) then
                        Replace the old individual
                    end if
                end if
            end for
            Generation counter cumulative gen = gen + 1
        end while
        SA algorithm operation
    end while
    Calculate the objective function value newObjV of the best initial cluster center
    Output: The objective function value Jb, cluster center
end do
```

Through the above procedure, the clustering centers of all types of data can be obtained. Then, the clustering centers were used as a standard for a class of data, while the remaining data are clustered.

V. EXPERIMENTAL PROCEDURE

The experiments were performed in the Key Laboratory of Fault Diagnosis of Petrochemical Equipment, Guangdong Province, China, using a fault diagnosis test platform of a large-scale petrochemical multistage centrifugal blower. The experimental platform was composed of 1) electric motor, 2) gearbox, 3) base platform, 4) coupling, 5) oil pipe, 6) fan, as shown in Fig. 8. The experimental platform was used to simulate a multistage centrifugal blower and common fault conditions of rolling bearings.

![FIGURE 8. Fault diagnosis test platform of large-scale rotating equipment used in the petrochemical industry.](image)

A. EXPERIMENTAL PROCEDURE

The experiments were performed on the rotary unit at a frequency of 20kHz and a rotational speed of 800r/min. The EMT390 vibrometer was used to collect signals under four different conditions, including three fault states (outer-ring wear, inner-ring wear, and bearing with one ball less), and the normal state (normal bearing). Table 1 lists the machine operation status types and the number of each corresponding datasets.

In this study, in order to clearly observe the effect of clustering in the two-dimensional plane, the mutual dimensionless indicators were standardized, principal component analysis was performed, and then the first and second principal component characteristic indicators were selected as input values. The data processing process can be seen in Fig. 9 and the relevant theory has described in detail in the literature [21].

Following main component analysis, Tables 2 to 5 demonstrate the corresponding contribution rate of each primary component between different types of data combinations respectively. When the first and second primary components were extracted, the contribution value of the selected sample was as high as 98.84%, 99.2%, 99.7%, and 99.2%. When the variance contribution rate of the main component exceeds...
B. SIMULATION AND RESULTS

In this study, different types of data combinations above were experimentally simulated. The characteristic indexes of the first and second principal components were the horizontal and vertical axis, respectively [38]. The experimental results are shown in Figs. 10 to 13.

Fig. 10 shows the results before and after clustering of a normal bearing and bearing with inner-ring wear. It was found that the true clustering number of the normal bearing was 49, the true clustering number of bearing with inner-ring wear was 49, and the total clustering accuracy was 100%. Fig. 11 shows the result of a normal bearing and bearing with outer-ring wear before and after clustering. The true clustering number of the normal bearing was 49, the true clustering number of that with outer-ring wear was 47, and the total clustering accuracy was 98%. According to Figs. 10 and 11, normal bearings and bearings with inner-ring and outer-ring wear can cluster well. In Fig. 12, the situation was different. The number of true judgments of the normal bearing was 42,

95%, the indicator can essentially represent the original characteristic information [37].

**TABLE 2.** Principal component analysis of normal bearing and bearing with one ball less.

| Component | Characteristic value | Contribution to variance | Cumulative |
|-----------|----------------------|--------------------------|------------|
| 1         | 3.977                | 0.7954                   | 0.7954     |
| 2         | 0.943                | 0.1886                   | 0.984      |
| 3         | 0.078                | 0.0156                   | 0.9996     |
| 4         | 0.002                | 0.0004                   | 1          |
| 5         | 0                    | 0                        | 1          |

**TABLE 3.** Principal component analysis of normal bearing and outer-ring wear.

| Component | Characteristic value | Contribution to variance | Cumulative |
|-----------|----------------------|--------------------------|------------|
| 1         | 4.487                | 0.8974                   | 0.8974     |
| 2         | 0.473                | 0.0946                   | 0.992      |
| 3         | 0.038                | 0.0076                   | 0.9996     |
| 4         | 0.002                | 0.0004                   | 1          |
| 5         | 0                    | 0                        | 1          |

**TABLE 4.** Principal component analysis of normal bearing and inner-ring wear.

| Component | Characteristic value | Contribution to variance | Cumulative |
|-----------|----------------------|--------------------------|------------|
| 1         | 4.840                | 0.968                    | 0.968      |
| 2         | 0.145                | 0.029                    | 0.997      |
| 3         | 0.015                | 0.003                    | 0.9999     |
| 4         | 0.001                | 0.0002                   | 1          |
| 5         | 0                    | 0                        | 1          |

**TABLE 5.** Principal component analysis of normal bearing, bearing with one ball less, outer-ring wear, and inner-ring wear.

| Component | Characteristic value | Contribution to variance | Cumulative |
|-----------|----------------------|--------------------------|------------|
| 1         | 4.605                | 0.921                    | 0.921      |
| 2         | 0.356                | 0.0712                   | 0.9922     |
| 3         | 0.037                | 0.0074                   | 0.9996     |
| 4         | 0.002                | 0.0004                   | 1          |
| 5         | 0                    | 0                        | 1          |
the number of true judgments of the bearing with one ball less was 33, and the total accuracy rate was 77%. In Fig. 13, there were three fault types, the correct number of judgments was 30, 33,37, respectively, and the accuracy rate was 73%, which was lower than the previous one. It can be seen that when the fault types are increased, the optimized clustering algorithm has a certain difficulty in resolution and the clustering accuracy rate decreases.

C. EXPERIMENTAL COMPARISON

In order to investigate the applicability and feasibility of the proposed algorithm for bearing fault diagnosis in engineering applications, the FCM and the proposed GA-SA-FCM algorithms were compared. In this comparative experiment, six sets of fault data were used. Each group and the corresponding fault types are listed in Table 6 and the experimental results are shown Table 7.

In Table 7, D method represents that the data processed by dimensionless method, and M-D method represents mutual dimensionless index. The samples of each fault combination are selected randomly, that is to say, in Table 7, samples in each row is the same, and different in different rows. Therefore, each experiment will show different results. Besides, to make the result more convincing, we repeat the experience for 20 times and average them.

Comparing the mutual dimensionless index and dimensionless index, we can find that the mutual dimensionless index show good performance in clustering. For FCM, the difference in accuracy was as high as 8.36%. For GA-SA-FCM the difference in accuracy was as high as 9.22%.

In addition, the accuracy of GA-SA-FCM is higher than the FCM. As is shown in Table 7, in No.3, the difference in accuracy was as high as 5.5%. Therefore, taking the accuracy after clustering as the fitness value function of the GA can improve the clustering effect, which is an improvement to the FCM algorithms.
Besides, we compared the proposed GA-SA-FCM algorithm with method proposed in letter [8] denoted as DSDF for briefly. Five types of data sets, including inner-ring wear, outer-ring wear, large gears teeth deficiency, large gears teeth deficiency and inner-ring wear combination, large gears teeth deficiency and outer-ring wear combination were compared. The comparison result were shown in Table 8 and 9.

In Table 8, the fault diagnosis accuracy rate of DSDF is 33.33%, and the GA-SA-FCM model is 48.89%, which means that the proposed algorithm is obviously better than that proposed by reference [8].

VI. DISCUSSION

In this paper, the accuracy of clustering and the distance value were taken as the objective function of SA-GA, which can improve the effect of clustering to a certain extent for the reason that the proposed method was able to solve the problem that cluster centers are easy influence by the outliers and local minimum problem.

In addition, the experience result shows that the mutual dimensionless indexes are more suitable for clustering algorithms than dimensionless one. for the reason that mutual dimensionless indicators are able to narrow the dimensionless the internal structure of the target distance, and then reduce the overlap of the same dimensionless index. Table 7 also shows that the advantage of our algorithm is obvious when the fault data is processed by mutual dimensionless.

However, compared with the FCM algorithm, the proposed method has high complexity. For the reason that the our algorithm need to calculate the accuracy, and then use the
SA for overcoming the local minimum problem of SA. As is shown in Table 7, with the increasing of the sample size the CPU time of our method is increasing rapidly than FCM. In addition, at present, we have no a systematic method to determine the weight of the accuracy of clustering and the distance value which have influence on the performance of clustering.

VII. CONCLUSION
It is generally believed that the clustering algorithm quality depends on the sum of the distance between the clustering center and the data points. It is also believed that the smaller the sum of distances, the better the clustering effect, while in fact, this is not the case. Therefore, the clustering distance and the clustering accuracy were used as the objective function, and the GA-SA algorithm was used for optimization. The experimental results demonstrated that the proposed algorithm can be better than the general clustering algorithm or the fuzzy clustering algorithm, which take the sum of the clustering distance as the objective function. Besides, with the increase of data distribution complexity and outliers, the algorithm proposed in this paper will have better performance. However, as we discussing in previous section, the proposed method still requires further improvements. Therefore, our further research will focus on determining the weight for the accuracy and distance ($J_b$) value and optimizing our algorithm to reduce its CPU time.

DATA AVAILABILITY
The data used to support the findings of this study are available from the corresponding author upon request.

CONFLICTS OF INTEREST
The authors declare that there are no conflicts of interest regarding the publication of this paper.

REFERENCES
[1] Y. G. Lei, “Clustering algorithm-based fault diagnosis,” in Intelligent Fault Diagnosis and Remaining Useful Life Prediction of Rotating Machinery. Butterworth-Heinemann, 2017.
[2] L. Wei, X. Li, L. Gao, and Y. Zhang, “A new convolutional neural network-based data-driven fault diagnosis method,” IEEE Trans. Ind. Electron., vol. 65, no. 7, pp. 5990–5998, Jul. 2018.
[3] S. Wang, I. Selesnich, G. Cai, Y. Feng, X. Xiu, and X. Chen, “Nonconvex sparse regularization and convex optimization for bearing fault diagnosis,” IEEE Trans. Ind. Electron., vol. 65, no. 9, pp. 7332–7342, Sep. 2018.
[4] L. Wang, Q. H. Zhang, C. Y. Ma, G. W. Lei, and J. He, “Research summary of dimensionless indicators in fault diagnosis technology,” Int. J. Mach. Tools Manuf., vol. 2, no. 2, pp. 19–22, 2014.
[5] S. Dong, X. Xu, and J. Luo, “Mechanical fault diagnosis method based on LMD Shannon entropy and improved fuzzy C-means clustering,” Int. J. Acoust. Vib., vol. 22, no. 2, pp. 714–720, Jun. 2017.
[6] Y. Hu, S. Zhang, A. Jiang, L. Zhang, W. Jiang, and J. Li, “A new method of wind turbine bearing fault diagnosis based on multi-masking empirical mode decomposition and fuzzy C-Means clustering,” Chin. J. Mech. Eng., vol. 32, no. 1, Dec. 2019.
[7] M. Mojarad, S. Nejatian, H. Parvin, and M. Mohammadpour, “A fuzzy clustering ensemble based on cluster clustering and iterative fusion of base classifiers,” Int. J. Speech Technol., vol. 49, no. 7, pp. 2567–2581, Jul. 2019.
[8] J. Xiong, Q. Zhang, Z. Peng, G. Sun, and Y. Cai, “Double sample data fusion method based on combination rules,” IEEE Access, vol. 4, pp. 7487–7499, 2016.
[9] G. X. Sun, A. S. Qin, Q. H. Zhang, Q. Hu, and X. S. Si, “A compound fault integrated diagnosis method for rotating machinery base on dimensionless immune detector,” in Proc. Control Decis. Conf., pp. 4390–4394, 2013.
[10] L. Y. Bai, S. Y. Hu, and S. H. Liu, “A fuzzy clustering algorithm based on simulated annealing and genetic algorithm,” Comput. Eng. Appl., vol. 41, no. 9, pp. 56–58, 2005.
[11] K. Polat, “Similarity-based attribute weighting methods via clustering algorithms in the classification of imbalanced medical datasets,” Neural Comput. Appl., vol. 30, no. 1, pp. 1–27, 2018.
[12] H. M. Chen, X. W. Li, Y. B. Zhang, and N. Chen, “An AK-LDMeans algorithm based on image clustering,” Int. Soc. Opt. Photon., Shanghai, Peoples R China, Young Sci. Forum, vol. 10710, p. 10710R, 2018.
[13] Z. Zhang, Q. Yang, and D. An, “An improved K-means algorithm for reciprocating compressor fault diagnosis,” in Proc. Chin. Control Decis. Conf., Jun. 2018, pp. 276–281.
[14] Q. Z. Wang and X. X. Wang, “Application of fuzzy clustering in mechanical fault diagnosis,” Bearing, no. 10, pp. 35–38, 2008.
[15] S. Wu, Y. Pang, S. Shao, and K. Jiang, “Advanced fuzzy C-means algorithm based on local density and distance,” J. Shanghai Jiaotong Univ., vol. 23, no. 5, pp. 663–642, Oct. 2018.
[16] W. Zhang and F. Z. Pan, “Fuzzy clustering based on genetic algorithms,” J. Hubei Univ., vol. 24, no. 2, pp. 102–104, 2002.
[17] Y. Qin, S. Ding, L. Wang, and Y. Wang, “Research progress on semi-supervised clustering,” Cognit. Comput., vol. 11, no. 5, pp. 599–612, Oct. 2019.
[18] S. Li and H. X. Zhang, “Semi-supervised clustering algorithm based on classifier,” Comput. Eng. Appl., vol. 47, no. 3, pp. 132–134, 2011.
[19] Y. Qin and S. F. Ding, “Survey of semi-supervised clustering,” Comput. Sci., vol. 46, no. 9, pp. 15–21, 2019.
[20] J. Xiong, Q. Zhang, G. Sun, X. Zhu, M. Liu, and Z. Li, “An information fusion fault diagnosis method based on dimensionless indicators with static discounting factor and KNN,” IEEE Sensors J., vol. 16, no. 7, pp. 2060–2069, Apr. 2016.
[21] J. Xiong, Q. Zhang, J. Wan, L. Liang, P. Cheng, and Q. Liang, “Data fusion method based on mutual dimensionless,” IEEE/ASME Trans. Mechatron., vol. 23, no. 2, pp. 506–517, Apr. 2018.
[22] J. Nayak, B. Naik, and H. S. Behera, “Fuzzy C-means (FCM) clustering algorithm: A decade review from 2000 to 2014,” in Computational Intelligence in Data Mining, vol. 32. New Delhi, India: Springer, 2015, pp. 133–149.
[23] S.-C. Tsai, J.-Y. Jiang, C. Wu, and S.-J. Lee, “A fuzzy similarity-based approach for multi-label document classification,” in Proc. 2nd Int. Workshop Comput. Sci. Eng., vol. 2, 2009, pp. 59–63.
[24] L. Jiang and F. Ding, “Improvement of the fuzzy C-means clustering algorithm,” Comput. Digit. Eng., vol. 38, no. 2, pp. 4–6, 2010.
[25] J. H. Holland, “Adaptation in natural and artificial systems,” Quart. Rev. Biol., vol. 6, no. 2, pp. 126–137, 1975.
[26] Z. A. A. Alyasseri, A. T. Khader, M. A. Al-Betar, A. K. Abasi, and S. N. Makhadmeh, “EEG signals denoising using optimal wavelet transform hybridized with efficient Metaheuristic methods,” IEEE Access, vol. 8, pp. 10584–10605, 2020.
[27] Q. Hu, Q. H. Zhang, A. S. Qin, and G. X. Sun, “The application of genetic programming for fault classifying of rotating machinery,” Comput. Appl. Softw., vol. 30, no. 10, pp. 57–59, 2013.
[28] L. Yu, F. Shi, H. Wang, and F. Hu, MATLAB Intelligent Algorithm 30 Case Studies, 2nd ed. Beijing, China: Beihang Univ. Press, 2015, pp. 188-196.
[29] X. Xiong, Q. Zhang, J. Li, H. Zhu, and H. Li, “Combining the multi-genetic algorithm and support vector machine for fault diagnosis of bearings,” Shock Vib., vol. 2018, no. 10, pp. 1–13, Oct. 2018.
[30] M. Steinbrunn, G. Moerkotte, and A. Kemper, “Heuristic and randomized optimization for the the ordering problem,” Vldh J., vol. 6, no. 3, pp. 191–208, 1997.
[31] J. H. Zhang and G. Li, “Simulated annealing algorithm and the application for solving TSP” Mod. Electron. Technol., vol. 29, no. 22, pp. 157–158, 2006.
[32] X. Bai, “Application of genetic algorithm fusing simulated annealing in document clustering,” Comput. Eng. Appl., vol. 42, no. 23, pp. 144–148, 2006.
[33] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” Science, vol. 220, no. 4589, pp. 671–680, 1983.
[34] J. Yang, A. G. Song, P. W. Xiong, and L. Y. Zhang, “A Gaussian model based method for nuclear detection robots to locate radiation sources,” Chin. High Technol. Lett., vol. 23, no. 3, pp. 282-288, 2013.
[35] P. Y. Kokate, M. Das, and R. N. K. Loh, “Rule-based control of slowly varying systems using compensator segmentation determined by simulated annealing,” *IEEE Trans. Ind. Electron.*, vol. 39, no. 4, pp. 359–367, Aug. 1992.

[36] F. Jin, C. H. Sun, and M. Zhong, “Improvement of fitness function in genetic algorithm,” *Mach. Des. Manuf.*, vol. 3, pp. 218–219, Jan. 2010.

[37] J. H. Chen, M. Y. Xie, Y. X. Wang, Z. W. Zhang, R. H. Peng, and X. R. Wang, “Classification of panax quinquefolium L. samples by principal component analysis,” *Natural Product Res. Develop.*, vol. 18, no. 2, p. 193, 2006.

[38] J. Zhang, W. Song, B. Jiang, and M. Li, “Measurement of lumber moisture content based on PCA and GS-SVM,” *J. Forestry Res.*, vol. 29, no. 2, pp. 557–564, Mar. 2018.

**JIANBIN XIONG** (Member, IEEE) was born in Hunan, China, in 1976. He received the B.E., M.A., and Ph.D. degrees from the Guangdong University of Technology, China. From 2012 to 2018, he was with the Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis, Guangdong University of Petrochemical Technology of China, where he is currently an Associate Professor, the Director of Subject Pattern Recognition and Signal Processing, and the Vice Director of the School of Computer and Electronic Information. He was a Visiting Professor with the Faculty of Science and Technology, University of Macau, in 2015. He is also an Associate Professor with the School of Automation Guangdong Polytechnic Normal University, Guangzhou, China, and the Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis. He has published over 40 papers in related international conferences, journal, and books. His research interests include fault diagnosis, signal processing, image processing, information fusion, and computer applications. He has served as a Reviewer of over 20 journals.

**XI LIU** received the B.S. degree from the School of Mechanical and Electrical Engineering, Xidian University, in 2004, the M.S. degree in computer-aided technology from Northwestern Polytechnical University, in 2007, and the Ph.D. degree in information and communication engineering from Sun Yat-sen University, in 2019. His research interests include sensor networks, application of information technology, and other fields.

**HONGBIN ZHU** was born in Guangdong, China, in 1995. He is currently pursuing the degree with the Department of Automatic Control, School of Automation, Guangdong University of Technology, China. He is a good Programer in MATLAB and C language. His current research interests include support the vector machine and fault diagnosis.

**HAIYING LI** was born in Guangdong, China, in 1995. She is currently pursuing the bachelor’s degree with the Guangdong University of Petrochemical Technology, China. She is also working with the Guangdong University of Petrochemical Technology. She is a good Programer in MATLAB and java language. Her current research interests include fuzzy C-means clustering algorithm and fault diagnosis.

**QINGHUA ZHANG** (Member, IEEE) received the B.Sc. degree in electrical automation from Henan Polytechnic University, China, in 1985, and the M.Sc. degree in industrial automation and the Ph.D. degree in control theory and control engineering from the South China University of Technology, China, in 1995 and 2004, respectively. He is currently a Professor and the Dean of the Guangdong Province Petrochemical Equipment Fault Diagnosis (PEFD) Key Laboratory, Guangdong University of Petrochemical Technology, China. His research interests include condition monitoring and fault diagnosis of rotating machinery, intelligence control, and applications of intelligent algorithms.

---

**XINGTONG ZHU** (Member, IEEE) was born in Hainan, China, in 1974. He received the B.S. degree from the Hunan University of Science and Technology, Xiangtan, China, in 1998, and the M.S. degree from Sun Yat-sen University, Guangzhou, China, in 2006. He is currently an Associate Professor with the School of Computer, Guangdong University of Petrochemical Technology, Maoming, China. His research interests include mechanical fault diagnosis and machine learning.