Research Article

Research on Novel Bearing Fault Diagnosis Method Based on Improved Krill Herd Algorithm and Kernel Extreme Learning Machine

Zhijian Wang,1,2 Likang Zheng,3 Junyuan Wang,1 and Wenhua Du1

1School of Mechanical Engineering, North University of China, Taiyuan, Shanxi 030051, China
2School of Mechanical Engineering, Xi’an Jiaotong University, Xi’an, Shanxi 030619, China
3School of Energy and Power Engineering, North University of China, Taiyuan, Shanxi 030051, China

Correspondence should be addressed to Junyuan Wang; wangyi01161013@163.com

Received 10 June 2019; Revised 21 August 2019; Accepted 6 September 2019; Published 28 November 2019

Academic Editor: Danilo Comminiello

Copyright © 2019 Zhijian Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this paper, a novel bearing intelligent fault diagnosis method based on a novel krill herd algorithm (NKH) and kernel extreme learning machine (KELM) is proposed. Firstly, multiscale dispersion entropy (MDE) is used to extract fault features of bearings to obtain a set of fault feature vectors composed of dispersion entropy. Then, it is imported into the kernel extreme learning machine for fault diagnosis. But considering the kernel function parameters \( \sigma \) and the error penalty factor \( C \) will affect the classification accuracy of the kernel extreme learning machine, this paper uses the novel krill herd algorithm (NKH) for their optimization. The opposite populations are added to the NKH in the initialization of population to improve its speed and prevent local optimum, and during the period of looking for the optimal solution, the impulse operator is introduced to ensure it has enough impulse to rush out of the local optimal once into the local optimum. Finally, in order to verify the effectiveness of the proposed method, it was applied to the bearing fault experiment of Case Western Reserve University and XJTU-SY bearing data set. The results show that the proposed method not only has good fault diagnosis performance and generalization but also has fast convergence speed and does not easily fall into the local optimum. Therefore, this paper provides a method for fault diagnosis under different loads. Meanwhile, the new method (NKH-KELM) is compared and analyzed with other mainstream intelligent bearing fault diagnosis methods to verify the effectiveness and accuracy of the proposed method.

1. Introduction

The rolling bearing is an indispensable part of the most rotating mechanical equipment. The loss of life and property and damage to the environment caused by the failure of the rolling bearing are very serious. In order to avoid bearing failure as much as possible, we can improve the bearing design, study and apply better materials and new technologies in the design and manufacturing stage, strengthen the quality control measures in the production process, and improve the assembly level of the bearing. In addition, another way to prevent catastrophic events of the rotating machinery equipment is to monitor and diagnose the working state of the equipment in real time so as to achieve effective control of the equipment. Therefore, it is of great significance to monitor the working state of the bearing in real time.

In recent years, various novel methods have been widely used to solve practical engineering problems [1–5]. And learning these novel methods [6–10] provides many ideas for fault diagnosis. In the field of bearing fault diagnosis, novel intelligent fault diagnosis methods emerge one after another in recent years, namely, the method based on statistics having Pearson’s correlation coefficient (PCC) [11], the method based on signal processing having modified variable modal decomposition (MVMD) [12], improved ensemble local mean decomposition (IELMD) [13], maximum kurtosis spectral entropy deconvolution (MKSED) [14], regression residual signal based on improved intrinsic timescale decomposition [15], enhanced singular spectrum
decomposition (ESSD) [16], weighted cyclic harmonic-to-noise ratio [17], time-frequency analysis [18], multipoint optimal minimum entropy deconvolution adjusted (MOMEDA) [19], and so on [20]. In recent years, with the development of big data, machine learning methods and deep learning methods have been widely used to solve practical engineering problems [21–26]. Machine learning methods or deep learning methods were applied in the field of bearing fault diagnosis, including the support vector machine (SVM) [27, 28], BP neural network (BP) [29], deep convolutional transfer learning network (CNN) [30], and kernel extreme learning machine (ELM) [31, 32]. These methods are good for fault diagnosis in most cases, but some are more subjective in choosing parameters. For example, the method used in reference [31] is highly subjective in selecting input weights and biases. Therefore, it is necessary to introduce the parameter optimization algorithm in fault diagnosis.

In the era of big data, many engineering problems need feature extraction. For example, Yin et al. used multiobjective feature extraction optimization to detect M/OD HV1 damages [33]. And, in the bearing fault diagnosis, the first step is how to collect useful information from the fault bearing. At present, there are many practical engineering problems that take vibration information as useful information, and there are endless methods for collecting vibration information [34–38]. And most scholars also collect bearing vibration information in the bearing fault diagnosis [39]. Then, how to extract the characteristic information from the vibration signal is the primary problem of intelligent fault diagnosis based on machine learning. As a tool to describe the uncertainty of signals, entropy has been widely used in recent years to extract fault features of rolling bearings [40, 41], such as sample entropy (SampEn) [42] and permutation entropy (PerEn) [43]. However, SampEn and PerEn need to calculate the distance between the two embedded dimensions and the embedded vector of each sample separately, which greatly increases the time for extracting features. In addition, most useful information cannot be extracted from a single time scale when analyzing the time series. In engineering practice, the optimal time scale of the original signal is often unknown. In order to solve this problem, the multiscale operation is introduced into feature extraction. Multiscale allows entropy to be extended to multiple time scales to provide an additional perspective when the time scales are uncertain. Like other entropy measurement methods, multiscale entropy aims to evaluate the complexity of the time series. One of the main reasons for using multiscale entropy is that the relevant time scales in the time series are not known. Therefore, analyzing the problem over multiple time scales will obtain more information. Moreover, bearing fault diagnosis based on multiscale entropy has been widely used in the field of intelligent fault, such as multiscale fuzzy entropy (MFE) [44] and multiscale permutation entropy (MPE) [45]. Rostaghi and Azami have proposed dispersion entropy (DE) [46] and multiscale dispersion entropy (MDE) [47]. MDE does not need to sort the amplitude of each embedded vector nor does it need to calculate the distance between any two compound delay vectors with embedded dimensions m and m + 1, which makes DE and MDE faster than PerEn, PerEn, and MperEn, significantly. Moreover, MDE has obvious advantages in distinguishing different types of dynamic signals, so MDE is more suitable than other methods for the extraction of the bearing vibration signal.

The result of feature extraction of each sample forms a feature vector. How to classify the feature vector accurately is the most critical problem of intelligent fault diagnosis. Three common classification methods based on machine learning are the BP neural network (BP), support vector machine (SVM), and extreme learning machine (ELM). However, in the traditional BP neural network, a large part of the network training parameters need to be set artificially, which is very random. In addition, BP is adjusted by the gradient descent method, which not only is easy to get into the local optimal but also suffers slow convergence [48]. More importantly, this operation is likely to lead to training failure. On the contrary, because the network structure model is not single, the parameters of the neural network need to be constantly adjusted in each iteration, so the diagnosis efficiency is not high, and it will easily fall into the phenomenon of underlearning or overfitting [49]. The SVM has great advantages in solving binary classification problems with small sample size [50], but it is difficult to implement large-scale sample training and solve multiple classifications [51]. However, the fault diagnosis of the rolling bearing is often a multiclassification problem, which means it is difficult for an SVM to implement the fault diagnosis of the rolling bearing and the high-level accuracy rate is difficult to reach. The ELM is a novel algorithm based on the single-hidden-layer feedforward neural network, which has the advantages of the simple mathematical model, global optimal solution, fast learning speed, less parameter selection, and high generalization, which has been applied in many fields [52]. However, the ELM randomly generates input weights and hidden-layer thresholds, resulting in instability of the algorithm. In order to solve this problem, the kernel function is introduced into the basis of the ELM to obtain the kernel extreme learning machine (KELM). The KELM solves the problem of random initialization in the ELM algorithm, and the number of nodes in the hidden layer can be determined accordingly, instead of being specified manually. In addition, in solving some practical engineering problems, the KELM shows high classification accuracy, good promotion ability, and high robustness [53].

In the intelligent fault diagnosis, the parameter optimization of the diagnosis method is also a key point. For example, particle swarm optimization (PSO) was used to optimize the filtering coefficient of deconvolution [54] and ant colony algorithm [55]. Gandomi proposed a krill herd optimization algorithm (KH) to solve the optimization problem in 2012 [56], which is based on the behavior of constantly updating the position of krills because of the predation of krills in nature. This algorithm can adjust the participation amount of exploration and development by observing the progress of solving the problem through various steps dynamically. In addition, Gandomi also compared the KH with eight common optimization methods [56], and the final result showed the KH converges faster than other optimization algorithms and does not easily fall into the local optimal. So, the KH can effectively solve various optimization problems and is superior to other
optimization algorithms. In addition, the online optimal method can also be used to solve such problems, and the research on online optimal methods is endless. For example, Yin et al. proposed the multivariate extremum seeking approach with the Newton method [57] and the multivariate fractional-order gradient-based extremum seeking approach [58], and the authors also verified their feasibility and advantages. Therefore, the online optimal method will be an important direction of parameter optimization in the future.

In the above context, the proposed optimization algorithms belong to the optimization algorithm of the population class but generally belong to the optimal problem [59]. That is to say, when an individual is in an optimal position, but other individuals cannot know whether this position is the local optimal or the global optimal, then all other individuals will move towards this optimal position and the whole population will fall into the local optimal eventually. At present, in order to solve this problem, the inertia weight of some individual movements can be increased so as to jump out of the local optimal through the high weight motion. But increasing the weight of a motion only works in one individual or one position adjustment, and it not only does not work in other positions but also may be the reason for falling into the local optimal. When an individual falls into the optimal position, the gradient of its position function changes greatly. If the individual moves than it normally does at the next position more, it is still near the optimal position after the move. In other words, each individual has an impulse associated with the previous step, which means the impulse is large when the previous step is large or the impulse is small when the previous step is small. However, the individual trapped in the optimal still does not rush out of the current optimal state after a certain number of moves, so it can be considered that this position is the global optimal position. In addition, it is also an effective method to improve the optimization algorithm by improving the initialization of the population. The OBL method proposed by Tizhoosh is one of the most effective methods to improve the initialization of population [60]. This algorithm increases the diversity of the search range and global search ability by adding the opposite population in the initial population, which avoids the whole population falling into the local optimal from the origin. When the whole population is searching for the global optimal, if the current optimal individual falls into the local optimal, then it may mislead other individuals to also enter the local optimal. However, the opposite of the individual does not usually go with it, but away from it. Therefore, the OBL method can effectively prevent the whole population from falling into the local optimum to some extent.

Under the same working conditions and the same bearing type, the vibration signals between different fault types are different because of different bearing fault types and fault diameters [61]. In addition, because the gearbox load is often changing according to the actual work requirements, the bearing of the axial and radial loads is constantly changing in practice. Too large or too small axial and radial loads may aggravate the vibration of the bearing [62] or increase the noise, or even produce screaming [63], which may lead to large fluctuation of the vibration signal and further affect the classification performance of the algorithm. In other words, an algorithm may only work properly under one load and not under other loads. In the actual gearbox fault diagnosis, in order to detect the health condition of a gearbox bearing, the load is deliberately set to a certain value, that is to say, to stop the normal operation of the gearbox to detect the health condition of the bearing, which will greatly reduce the economic benefit of the gearbox. Therefore, it is of great significance to consider various load conditions when studying intelligent fault diagnosis of bearings.

Based on the above, this paper proposed a novel bearing intelligent fault diagnosis method based on the kernel extreme learning machine (KELM). A novel krill herd algorithm (NKH) is used to optimize kernel function parameters \( \sigma \) and the error penalty factor \( C \) in the KELM. Opposite-based learning (OBL) and impulse operator are introduced in the optimization algorithm to improve the global search ability of the individual krill and prevent the krill group from falling into the local optimal and increase the robustness of the algorithm. Firstly, in order to verify the correctness of the algorithm proposed in this paper, the bearing data set of Case Western Reserve University is used for this experiment, and the influence of different loads on the experimental results is considered in this experiment. Then, in order to test the performance of the intelligent fault diagnosis method proposed in this paper, it is compared with other methods based on machine learning.

2. Feature Extraction

2.1. Multiscale Operation. In the analysis of fault vibration signals, a more appropriate choice of the time scale usually means that more useful information can be obtained from the original signal. However, in the engineering practice, the best time scale of the original signal is often unknown. In order to solve this problem, the signal of multiple time scales should be considered comprehensively.

Multiscale allows entropy to be extended to multiple time scales to provide an additional perspective when time scales are uncertain. Like other entropy measurement methods, multiscale entropy aims to evaluate the complexity of the time series. Therefore, analyzing the problem through multiple time scales will obtain more information. For a given time series \( x = \{x_1, x_2, x_3, \ldots, x_N\} \) whose length is \( N \), the conversion equation of the multiscale operation is as follows:

\[
u_j^{(\tau)} = \frac{1}{\tau} \sum_{b=\lfloor j-1 \rfloor}^{\lfloor j \tau \rfloor} x_b, \quad 1 \leq j \leq \left[ \frac{N}{\tau} \right] = N, \tag{1}
\]

where \( u_j^{(\tau)} \) represents the \( j \)th element in the sequence \( u \) with the time scale of \( \tau \), \( N \) represents the length of the sequence \( u \), and \( x_b \) represents the \( b \)th element in the sequence \( x \).

A multiscale coarse granulation operation with a time scale of 2 and 3, respectively, is shown in Figures 1 and 2.

2.2. Dispersion Entropy (DE). Dispersion entropy can be used to express the dispersion degree of a time series. For a given time series \( x = \{x_1, x_2, x_3, \ldots, x_N\} \) whose length is \( N \), the DE consists of the following four steps:
where \( \sigma \) is the standard deviation of the sequence and \( \mu \) is the mean of the sequence.

Secondly, after obtaining the sequence \( y \), it is linearly mapped to \( z \), and the equation is as follows:

\[
 z^c_j = \text{round}(c \times y_j + 0.5),
\]  

where \( c \) represents the number of types of classifications, \( z^c_j \) represents the \( j \)th element of the sequence, and \( \text{round}(\cdot) \) represents the rounding operation.

(2) The equation of the embedding vector \( z_{i}^{m,c} \) is as follows:

\[
 z_{i}^{m,c} = \{z_{i+1}^{c+1}, z_{i+2d}^{c}, \ldots, z_{i+(m-1)d}^{c}\},
\]

where \( i = 1, 2, 3, \ldots, N - (m - 1)d \), \( m \) is the embedding dimension, and \( d \) is the time delay.

Then, each embedded vector is mapped to the dispersion pattern; the mapping relationship is as follows:

\[
 z_{i}^{c} = v_{0}, v_{1d}, v_{2d}, \ldots, v_{i(m-1)d} = v_{m-1},
\]

So there are \( c^m \) different dispersion patterns.

(3) For each possible dispersion pattern, the equation of its relative probability is

\[
 p\left(\pi_{i_0, i_1, \ldots, i_{m-1}}\right) = \frac{\text{number}\{ i \mid i \leq N - (m - 1)d, z_{i}^{m,c} \text{ has type } \pi_{i_0, i_1, \ldots, i_{m-1}}\}}{N - (m - 1)d},
\]  

where the possibilities of the dispersion pattern are \( 3^2 = 9 \), that is,

\[
\pi_{0,1,...,m-1} = \{\pi_{11}, \pi_{12}, \pi_{33}, \pi_{21}, \pi_{22}, \pi_{33}, \pi_{32}, \pi_{23}\}. \]

Its detailed equation is as follows:

\[
\begin{align*}
 z_{1}^{2,3} &= \{3, 3\} \quad \rightarrow \quad \pi_{33}, \\
 z_{2}^{2,3} &= \{3, 1\} \quad \rightarrow \quad \pi_{31}, \\
 z_{3}^{2,3} &= \{1, 3\} \quad \rightarrow \quad \pi_{13}, \\
 z_{4}^{2,3} &= \{3, 2\} \quad \rightarrow \quad \pi_{32}, \\
 z_{5}^{2,3} &= \{2, 1\} \quad \rightarrow \quad \pi_{21}, \\
 z_{6}^{2,3} &= \{1, 1\} \quad \rightarrow \quad \pi_{11}, \\
 z_{7}^{2,3} &= \{1, 3\} \quad \rightarrow \quad \pi_{13}, \\
 z_{8}^{2,3} &= \{3, 2\} \quad \rightarrow \quad \pi_{32}, \\
 z_{9}^{2,3} &= \{2, 2\} \quad \rightarrow \quad \pi_{22}, \\
 z_{10}^{2,3} &= \{2, 1\} \quad \rightarrow \quad \pi_{21}, \\
 z_{11}^{2,3} &= \{1, 3\} \quad \rightarrow \quad \pi_{13}.
\end{align*}
\]
So the probability of each dispersion pattern is as follows:

\[
\begin{align*}
\ p(\pi_{11}) &= \frac{1}{11}, \\
\ p(\pi_{12}) &= \frac{0}{11}, \\
\ p(\pi_{13}) &= \frac{3}{11}, \\
\ p(\pi_{21}) &= \frac{2}{11}, \\
\ p(\pi_{22}) &= \frac{1}{11}, \\
\ p(\pi_{23}) &= \frac{0}{11}, \\
\ p(\pi_{31}) &= \frac{1}{11}, \\
\ p(\pi_{32}) &= \frac{2}{11}, \\
\ p(\pi_{33}) &= \frac{1}{11}
\end{align*}
\] \hspace{1cm} (10)

Its dispersion entropy is calculated as 

\[
DE(x, m, c, d) = DE(x_1, 2, 3, 1) = 1.8642
\]

and normalized as 

\[
NDE = \frac{DE}{\ln(3^2)} = 0.85.
\]

2.3. Multiscale Dispersion Entropy (MDE). When dealing with the fault of the bearing vibration signal, it is often difficult to know the most suitable time scales, and signal nonstationarity and irregularity tend to be very strong; in order to solve this problem well, Azami et al. proposed multiscale discrete entropy (MDE) in 2017 [47] and proved that the MDE approach based on processing of the nonstationary signal has a strong ability of feature extraction. The flow of MDE is shown in Figure 3.

\[
x = [9, 8, 7.2, 5.3, 6.4, 3.2, 0.89, 10.6, 5.2, 4.3, 6.4, 7.8].
\] \hspace{1cm} (11)

Here is a simple example: supposing there is a set of time series, the embedding dimension \(m\) is 2, the time delay \(d\) is 1, the types of classifications \(c\) are 3, and the time scale is 3. Firstly, the coarse granulation of the sequence is carried out according to the time scale. And the time scale 2 of the sequence \(y^{(2)} = [8.5, 6.25, 4.8, 5.745, 4.75, 7.2]\) and the time scale 3 of the sequence \(y^{(3)} = [8.07, 4.97, 5.56, 6.17]\) are obtained. Then, the dispersion entropy of each sequence is calculated to get MDE = \(\{DE^{(1)}, DE^{(2)}, DE^{(3)}\} = [2.0198, 1.0549, 1.0986]\). Finally, it is normalized to NMDE = \(\{0.9193, 0.4801, 0.5\}\).

3. The Proposal of NKH-KELM

3.1. Novel Krill Herd Algorithm (NKH). Gandomi proposed the krill herd algorithm (KH) in 2012 by studying krill herd activity rules [56]. In this algorithm, the solution of the optimization problem is expressed by the position state of the krill, and the optimal solution is constantly sought by the change of the individual position in the process of krill foraging. The Lagrange model is used to describe the \(K\)th location change of the \(i\)th krill in a group of krills including \(L\) krills during the foraging process:

\[
\frac{dX_i}{dt} = N_i + F_i + D_i,
\] \hspace{1cm} (12)

where \(N_i\) is the induced movement among krills, \(F_i\) is the foraging movement among krills, and \(D_i\) is the random disturbance of krills.

The induced motion velocity \(N_i^{(k)}\) of the \(i\)th krill is expressed as follows:

\[
N_i^{(k)} = N_{\text{max}}a_i^{(k)} + \omega_n N_i^{(k-1)},
\]

\[
a_i^{(k)} = a_i^{(k)}\text{local} + a_i^{(k)}\text{target},
\]

\[
a_i^{(k)}\text{local} = \sum_{j=1}^{L} K_{ij}X_{ij},
\] \hspace{1cm} (13)

\[
X_{ij} = \frac{X_i - X_j}{||X_i - X_j|| + \epsilon},
\]

\[
K_{ij} = \frac{K_i - K_j}{K_\text{worst} - K_\text{best}},
\]

where \(N_{\text{max}}\) is the maximum induced velocity; \(\omega_n\) is the inertia weight (\(\omega_n \in [0, 1]\)); \(a_i^{(k)}\text{local}\) is the local impact from the neighbor krill; \(a_i^{(k)}\text{target}\) is the current local influence from the optimal krill; \(K_i\) and \(K_j\) are the fitness function values of krills \(i\) and \(j\), respectively; \(K_\text{worst}\) and \(K_\text{best}\) are the worst and the best fitness function value of the current krill, respectively; \(K_{ij}\) is the fitness function value of the \(i\)th krill related to the \(j\)th krill; \(X_{ij}\) is the position of the \(i\)th krill related to the \(j\)th krill; and \(\epsilon\) is a normal number which avoided the singularity of equation (18).
The foraging movement $F_i^{(k)}$ of the $i$th krill is represented by the following equation:

$$F_i^{(k)} = v_f \left( \beta_f^{(k)} + \beta_b^{(k)} \right) + \omega_f F_i^{(k-1)},$$  

(14)

where $v_f$ is the speed of foraging; $F_i^{(k-1)}$ is the direction of the last foraging; $\beta_f^{(k)}$ is the attraction of food to the $i$th krill; $\beta_b^{(k)}$ is the influence of the optimal krill on the $i$th krill until the current moment; and $\omega_f$ is the inertia weight ($\omega_f \in [0, 1]$).

The random disturbance $D_i^{(k)}$ of the $i$th krill is expressed by the following equation:

$$D_i^{(k)} = D_{\max} \delta,$$

(15)

where $D_{\max}$ is the maximum velocity of random diffusion and $\delta$ is the direction vector of random diffusion disturbance ($\delta \in [-1, 1]$).

In order to improve the global search ability and convergence speed of the krill herd algorithm, the opposite population in the initialization of the krill herd is added. Assuming that the position of the krill is $X_i = (c_1, c_2, c_3, \ldots, c_D)$ in the $D$-dimensional space, where $c_j \in [c_{\min}(j), c_{\max}(j)], j \in [1, D]$, the position equation of the opposite krill is

$$X_i^{\ker} = (c_1, c_2, c_3, \ldots, c_D),$$

$$c_j^{\ker} = (c_{\min}(j) + c_{\max}(j) - c_j).$$

(16)

Through opposite-based learning (OBL), the diversity of the krill group is increased so as to increase the exploration scope and robustness of the krill group.

There may not only be one optimal solution in the $D$-dimensional space that describes the krill herd algorithm; that is to say, there are many local solutions. The impulse operator is introduced to make a krill whose previous movement is very large have a high velocity in the next movement so as to rush out of the local optimal solution. So equation (8) can be improved to get the following equation:

$$\frac{dX_i^{\ker \text{new}}}{dt} = N_i + F_i + D_i + \rho \frac{d^2X_i^{\ker \text{old}}}{dt^2},$$

(17)

where $\rho$ is the impulse coefficient and $(d^2X_i^{\ker \text{old}})/(dt^2)$ is the acceleration of the $i$th krill in the last movement. The flow chart of the NKH is shown in Figure 4.

3.2. The Kernel Extreme Learning Machine Optimized by Novel Krill Herd Algorithm (NKH-KELM). Consider a sample data set $(x_i, t_i)$ whose capacity is $N$, where $x_i = [x_{i1}, x_{i2}, x_{i3}, \ldots, x_{iN}]^T$ is the sample input data, $N$ is the length of each vector $x_i$, and $t_i = [t_1, t_2, t_3, \ldots, t_N]^T$ is the sample output value. For a single-layer forward neural network (SLFN) with $L$ nodes of the hidden layer, its flow is shown in Figure 5.

In Figure 5, $h(a_i x_i + b_i)$ is the activation function, $a_i$ is the weight vector between the $i$th node of the hidden layer and the node of the input layer, and $b_i$ is the bias of the $i$th hidden layer, $\beta_i$ is the weight vector between the $i$th node of the hidden layer and the output layer, and $y_i$ is the output value. So the output equation is

$$y_i = \sum_{i=1}^{L} \beta_i h(a_i x_i + b_i), \quad i = 1, 2, 3, \ldots, N.$$  

(18)

When the actual output of the extreme learning machine can approximate the expected output infinitely, in other words, $\sum_{i=1}^{N} \|y_i - t_i\| = 0$,

$$t_i = \sum_{i=1}^{L} \beta_i h(a_i x_i + b_i), \quad i = 1, 2, 3, \ldots, N.$$  

(19)

The output of the hidden layer is represented by a matrix

$$H = \begin{bmatrix} h(a_1 x_1 + b_1) & \ldots & h(a_N x_1 + b_1) \\ \vdots & \ddots & \vdots \\ h(a_1 x_1 + b_1) & \ldots & h(a_N x_1 + b_1) \end{bmatrix},$$

and $T$ is the expected output vectors. The least-square method is used to determine the output weight vector $\beta_i$ of the ELM. Its equation is as follows:

$$\beta = H^+ T = H^T \left( \frac{I}{C} + HH^T \right)^{-1} T,$$

(20)

where $H^+$ is the Moore–Penrose generalized inverse matrix of the hidden-layer output matrix, $C$ is the penalty factor, and $I$ is the identity matrix. In order to prevent the roots from deviating from zero, a constant matrix $I/C$ is added to the matrix $HH^T$ to improve the stability and generalization ability of the results.

In terms of classification, for any linear indivisible data set $X$ in a low-dimensional space, there is always a mapping $k$ that maps the samples to a high-dimensional feature space $X \rightarrow k(x)$, which enables the data set to be linearly separated. However, the high-dimensional feature space will greatly increase the calculation amount, and more importantly, the generalization performance of the algorithm will decline with the increase of the dimension, which is very unfavorable to the intelligent fault diagnosis of bearings. In other words, an algorithm is only applicable to intelligent fault diagnosis of bearings under one working condition, but it is often not applicable to intelligent fault diagnosis of bearings under another working condition. The kernel extreme learning machine (KELM) is borrowed from the idea of the support vector machine (SVM) and adopts kernel function to replace the feature mapping of the ELM hidden-layer node, which can avoid the problem of the dimension disaster to a certain extent. Compared with the ELM, the KELM does not need to artificially determine the number of hidden-layer nodes and also avoids the randomization operation of input weight and input bias, thus improving the generalization ability and classification accuracy of the model. In the KELM, only the appropriate kernel function parameters need to be selected to obtain the output weight.

For the KELM, the output vector is expressed as follows:

$$y = \ker(x)H^T \left( \frac{I}{C} + HH^T \right)^{-1} T.$$  

(21)

The kernel function is expressed as follows:

$$\ker(x, x_j) = k(x_i)k(x_j).$$  

(22)

In the selection of kernel function, since there is no prior knowledge about the classification of fault bearings, and the
KELM is often expected to have a strong ability to explore the local optimum after a certain time of use, radial basis function (RBF) is selected as the kernel function of the KELM [64]. The equation of the Gaussian kernel function is as follows:

\[
\text{ker}(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right),
\]

where \(x_i\) is the input vector during training, \(x_j\) is the input vector during testing, and \(\sigma\) is the width parameter of the kernel function.

So the output equation of the KELM is

\[
y = \begin{bmatrix} \text{ker}(x_i, x_1) \\ \vdots \\ \text{ker}(x_i, x_L) \end{bmatrix}^T H^T \left(\frac{I}{C} + HH^T\right)^{-1} T,
\]

where \(L\) is the dimension of the input vector.

Considering that the kernel function parameters \(\sigma\) and the error penalty factor \(C\) in the KELM have a great influence on the results, the novel krill herd algorithm (NKH) is used to optimize the two parameters to improve the classification accuracy and robustness of the KELM. The generalized mean-square error GMSE can directly reflect the regression performance of the KELM [65]. The equation of the GMSE is

\[
\text{GMSE} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2,
\]

where \(n\) is the number of test samples and \(\hat{y}_i\) and \(y_i\) are the estimated and actual values of test samples, respectively.

In this paper, the KELM algorithm is used to solve the classification problem, and the accuracy of the algorithm classification can be verified by 10-fold cross-validation. In other words, the data set is divided into ten parts, and turns are taken to use nine of them as training data and the remaining one as test data, which can be used to evaluate the performance of parameters, reduce overfitting to a certain extent, and obtain as much effective information as possible from limited data to improve model performance [66].

Then, the GMSEs which are obtained ten times are averaged. Finally, the result is used as the fitness function \(K\):

\[
K = \text{GMSE} = \frac{1}{k} \sum_{i=1}^{10} \text{GMSE},
\]

where \(k\) is the number of experiments and is taken as 10.

From what has been discussed above, the flow chart of NKH-KELM is shown in Figure 6. As can be seen from Figure 6, the fault diagnosis method proposed in this paper consists of three parts, namely, fault feature extraction, NKH-KELM training, and NKH-KELM testing. In feature extraction, vibration signals are collected by accelerometers, sample design is carried out, and MDE feature extraction is finally carried out so as to extract useful information from the original vibration signals. In NKH-KELM training, the NKH with OBL and impulse operator is used to optimize the width parameters \(\sigma\) and penalty factors \(C\) of kernel function, and then the feature space extracted from the training set is imported into the KELM for training. In the testing of NKH-KELM, the feature space extracted from the feature set is imported into the classifier which is trained in the previous step so as to conduct fault diagnosis and verify the feasibility of the method proposed in this paper.
According to the OBL criterion, join the opposite krill

Training NKH-KELM

Testing NKH-KELM

Fault diagnosis result

Figure 6: Flow chart of NKH-KELM.
4. Bearing Intelligent Fault Diagnosis Based on NKH-KELM

4.1. Fault Diagnosis Experiment Based on Case Western Reserve University Bearing Data Set

4.1.1. Description of Data. In order to verify the feasibility of the proposed method, this paper adopts the test bearing data set provided by Case Western Reserve University to carry out the experiment. The program code of all algorithms was written in MATLAB R2016a, and then it was run on a computer with CPU i7-5500U@2.40 GHz, RAM 12.00 GB, and 64 bit Win10 operating system.

As shown in Figure 7, the bench consists of a 2 horsepower motor (left), a torque sensor/encoder (center), a dynamometer (right), and other control equipment (not shown in Figure 7). The motor shaft is supported by test bearings. The single point failure of all test bearings was caused by Electric Discharge Machining (EDM), and the fault diameters were 0.007 inch, 0.014 inch, and 0.021 inch, respectively. The driving-end bearing (this paper only considers the driving-end bearing data) adopts the 6205-2RS bearing of SKF company. All vibration data were collected by an accelerometer, which was placed at the drive end of the motor housing with the sampling frequency of 12 kHz.

The larger the fault size, the greater the bearing fault degree. 0.007 inch, 0.014 inch and 0.021 inch are considered in this paper. In addition, for the motor load, the working conditions with no load (0hp), small load (1hp), and large load (3hp) are considered. For the convenience of the following instructions, a notation is used to represent the tested data. For example, IR014_3 represents the bearing of an inner ring failure type with a load of 3hp and a failure diameter of 0.014 inch. Table 1 shows the specifications of test bearings.

As can be seen from Table 1, there are 30 kinds of test bearing data sets. Since each fault type corresponds to 3 fault diameters besides a healthy bearing, there are 10 kinds of bearings. For each test bearing data set, every 5000 sampling points are selected as one sample, so each data set has 24 samples and the whole data set can be divided into 720 samples. The time-domain graphs of a sample in the bearing data of each test under three loading conditions are shown in Figures 8–10, respectively.

From the above time-domain diagram, it can be clearly seen that, under the same load, the same fault type, and different fault diameters, the peak size and peak occurrence time of the time-domain diagram are significantly different. For example, when the load is 3 hp, there are three different time-domain diagrams corresponding to the ball fault. In addition, in the case of the same fault type and fault diameter under different loads, there are obvious differences in the peak size and peak occurrence time of time-domain graphs. Moreover, as can be seen from Table 2, the maximum value of the signal may increase or decrease with the increase of load. For example, in the rolling body fault with the fault diameter of 0.007 inch, with the increase of load, the maximum value of the signal is 3.4254, 3.0960, and 3.1301 respectively. Moreover, with the increase of load, the minimum value, peak value, and root-mean-square value all change to some extent. In addition, with the same load and fault type, but with different fault diameters, the peak values of time-domain graphs also have obvious differences. Therefore, this further proves the necessity of considering load and fault diameter in intelligent fault diagnosis.

4.1.2. Fault Bearing Classification in Case Western Reserve University Bearing Data Set Based on NKH-KELM. According to the intelligent fault diagnosis method proposed in this paper, firstly, MDE is used to extract multiscale features from the original vibration signals. For the choice of the time scale \( \tau \), if the time scale is too small, that is, the observation angle of the signal is too small, it is impossible to extract more useful information and affect the subsequent operations. Conversely, if the time scale \( \tau \) is too large, it will increase the memory requirements and computing time of
the computer. Secondly, for the selection of the embedding dimension $m$, if $m$ is too small, it is difficult to detect the tiny dynamic behavior of signals. Moreover, as can be seen from the time-domain graphs in the previous section, the tiny dynamic behavior of signals may bring about changes in the fault type or fault diameter. On the contrary, if the embedding dimension $m$ is too large, although it can obtain more reliable entropy results, it will greatly increase the memory requirements and computing time of the computer.

Therefore, according to the above contents, in order to extract more useful information on the limited computer memory, the time scale $\tau$ is assumed to be 20 and the embedding dimension $m$ to be 2 \cite{67}. In addition, when the time delay $d$ is greater than 1, aliasing may occur. Therefore, the time delay $d$ in this paper is taken as 1. Figure 11 shows the MDE extraction results of the health condition of the bearing in the test bearing data set under three different loads.

**Figure 8:** Time-domain graph of the original signal when the load is 0 hp.

**Figure 9:** Time-domain graph of the original signal when the load is 1 hp.
In this experiment, the extracted feature space is a matrix with a size of $720 \times 20$, in which 720 is the number of samples and the dimension of each feature vector is 20. Before using NKH-KELM for classification, the number of training samples is selected as 540, the number of test samples is selected as 180, and the classification label is selected as 4 (N, IR, B, and OR). Firstly, the training sample set and the novel krill herd algorithm are used to optimize the kernel function parameters $\sigma$ and the error penalty factor $C$ in the KELM. In the initialization of the krill group, the number of krills is selected to be 25, and the maximum allowable number of krill position updates is 200 [68]. After OBL, 25 opposition krills are produced, so the individual number of whole krill groups is 50. After introducing the impulse operator, the value of the impulse coefficient has a great influence on the robustness of the krill herd algorithm. If the impulse coefficient is too small, the krill may fall into the local optimal and is difficult to rush out of the local optimal because of the fact the krill’s impulse is too small. If the impulse coefficient is too large, then the impulse will be too large and the krill will always be out of the global optimal; that is to say, once the krill enters the global optimal, the krill herd algorithm will never reach the global optimal. When the impulse factor is greater than or equal to 1, the next move of the krill is likely to occur in a situation that will never be global optimal, which is the worst. Therefore, when discussing the selection of the impulse coefficient in this paper, it should first be set to be less than 1 and then its minimum value is set to be 0.1 and the interval between values of two adjacent groups is set to be 0.1. The experiment is repeated five times, and the results are shown in Figure 12.

It can be seen from Figure 12 that when the impulse coefficient $p$ is 1, the optimal fitness hardly changes after reaching $10^{-3}$. That is to say, the phenomenon that the impulse is too large to enter the optimal solution area and the next movement will rush out of the optimal solution area may happen, which indirectly proves the correctness of the
previous hypothesis. As can be seen from Figure 12, when
the impulse coefficient $p$ is 0.6 and krill moves 200 times, the
average fitness of the optimal krill individual is $2.165 \times 10^{-9}$,
the error penalty factor $C$ after optimization is 15, and the
width parameter $\sigma$ of the kernel function is 0.53. Secondly,
the optimized KELM is used to train and test the feature
space extracted from MDE. And the input of the KELM is
composed of 720 samples of MDE, of which 540 are used for
training and the remaining 180 samples are used for testing.
Then, the process of training is according to the KELM’s
kernel mapping rules. Finally, the output of the KELM is
obtained, that is, the prediction labels of the input sample.
They were compared with the actual labels of samples, and
the KELM was trained based on this basis. The experiment is
repeated five times. The average fault diagnosis accuracy is
shown in Table 3, and the fault diagnosis of the training and
test of a certain experiment is shown in Figure 13. In Figure
13, the abcissa is the number of sample sets, the ordinate
is the state of the sample, the circle represents the actual label
of the sample, and the asterisk represents the prediction label
of the sample. As can be seen from Figure 13, the classification
accuracy of the training process and test process
reached 99.2593% and 98.3333%, respectively. In addition,
the correct rate reached 100% in the diagnosis of the normal
bearing set, and the other three kinds of bearings all had a
few cases of wrong diagnosis. This also explains the feasibil-
ity of the proposed method. Moreover, the sample of error
diagnosis mainly focuses on the rolling body, which may be
related to the working environment of the rolling body and
other factors.

Figure 11: Dispersion entropy of each scale of the vibration signal of the test bearing data set under three kinds of loads. (a) Load = 0 hp. (b) Load = 1 hp. (c) Load = 3 hp.
In order to objectively test the performance of the proposed method, the same data set is imported into other algorithms, such as the K-nearest neighbor algorithm (KNN) [69], naive Bayes classification algorithm (NB) [70], symbol dynamic entropy + support vector machine (SVM) [27], radial basis network (RBF) [71], extreme learning machine (ELM) [31], and multilayer extreme learning machine [32]. The experiment is repeated five times, and the average value of the five times is taken as the evaluation index. The results are shown in Table 4 and Figures 14 and 15.

It can be seen from Table 4 and Figure 14 that the performance of the kernel extreme learning machine

<table>
<thead>
<tr>
<th>Types of bearings</th>
<th>Average accuracy of the training sample set (%)</th>
<th>Average accuracy of the test sample set (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Inner ring fault</td>
<td>98.9</td>
<td>97.6</td>
</tr>
<tr>
<td>Ball fault</td>
<td>99.8</td>
<td>97.1</td>
</tr>
<tr>
<td>Outer ring fault</td>
<td>97.9</td>
<td>98.4</td>
</tr>
<tr>
<td>Average</td>
<td>99.1</td>
<td>98.3</td>
</tr>
</tbody>
</table>
optimized by the krill herd algorithm which is improved by
OBL and impulse operator is better than that of the KELM
optimized by the standard krill herd algorithm, which also
proves the superiority of the novel krill herd algorithm
proposed in this paper and the correctness of the algorithm
introduced by OBL and impulse operator. Figure 15 shows
the average value of the confounding matrix of the six al-
gorithms in the five experiments, where the horizontal
coordinate is a certain prediction label, the vertical co-
ordinate is the actual label of the sample, and the number in
the grid represents the proportion of the prediction label in
the classification and the actual label. It can be seen from
Figure 15 and Table 4 that the seven methods used for
comparison and the method proposed in this paper reach
100% in the classification of healthy bearings. The
phenomenon may be that vibration signals of the health bearing
and the vibration signal of the fault bearing are different
greatly, the extraction of the vibration signal level of Case
Western Reserve University is high, or the feature extraction
method (MDE) is very superior, which leads to all kinds of
algorithms on the healthy bearing reach the classification
accuracy of 100%. Among the three fault types, NKH-KELM
is superior to the other seven algorithms, which proves that
NKH-KELM has certain advantages of fault diagnosis based
on machine learning. In addition, in the fault diagnosis of
the ball, the six different algorithms used for comparison
have not more than 95%, among which NB has only 49%. It
can also be seen from Figure 15 that the six algorithms
overlap greatly in the classification of ball faults and outer
ring faults. In other words, the ability of these six algorithms
to distinguish the ball fault and outer ring fault is not high.
However, the ball fault diagnosis accuracy of NKH-KELM
reaches 97.1%, and the outer fault diagnosis accuracy reaches
98.4%, which indicates that NKH-KELM has certain ad-
vantages over other algorithms in distinguishing ball faults
and outer ring faults.

4.2. Fault Diagnosis Experiment Based on XJTU-SY Bearing
Data Set

4.2.1. Description of Data. As shown in Figure 16, the
bearing test is composed of an AC motor, a motor speed
controller, a supporting shaft, a hydraulic loading system,
etc. [72]. The radial force is generated by the hydraulic
loading system and applied to the bearing housing under the
test. The speed is set and maintained by the AC induction
motor speed controller. In this experiment, the bearing
model is LDK UER204, and its specifications are shown in
Table 5.

4.2.2. Fault Bearing Classification in XJTU-SY Bearing Data
Set Based on NKH-KELM. In order to test the generalization
of the method proposed in this paper, the XJTU-SY bearing
data set with the sampling frequency of 37.5 Hz was selected.
The last 5000 points in the first 108 files of Bearing2_1 (inner
race fault), Bearing2_2 (outer race fault), and Bearing2_3
(cage fault) are selected as failure samples, and the first 5000
points in the first 36 samples of Bearing3_4 are taken as
healthy samples. In other words, the input of this experiment

<table>
<thead>
<tr>
<th>Feature extraction</th>
<th>Classification method</th>
<th>Average classification accuracy (%)</th>
<th>Average computing time (s)</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDE</td>
<td>KELM + standard KH</td>
<td>88.3</td>
<td>0.6398</td>
<td>0.1059</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>96.5</td>
<td>0.5398</td>
<td>0.1447</td>
</tr>
<tr>
<td></td>
<td>NB</td>
<td>77.8</td>
<td>1.1771</td>
<td>0.5369</td>
</tr>
<tr>
<td></td>
<td>RBF</td>
<td>94.8</td>
<td>3.1687</td>
<td>0.3681</td>
</tr>
<tr>
<td>SDE</td>
<td>SVM</td>
<td>95.5</td>
<td>3.3622</td>
<td>0.4369</td>
</tr>
<tr>
<td>mWAE</td>
<td>ELM</td>
<td>95.2</td>
<td>0.7469</td>
<td>0.2173</td>
</tr>
<tr>
<td>MPE</td>
<td>MELM</td>
<td>97.4</td>
<td>0.8312</td>
<td>0.0871</td>
</tr>
</tbody>
</table>
is 360 samples, including 36 healthy samples and 324 failure samples. The parameter selection of feature extraction is the same as that of the previous experiment. Then, the feature space after feature extraction is a matrix of $360 \times 20$. The ratio of the training sample set and the test sample set is 3 to 1. In other words, the feature space of the training sample set is $270 \times 20$, and that of the test sample set is $90 \times 20$. Then, the training sample set was imported into NKH-KELM with an impulse coefficient of 0.6 to adjust the model for training. Finally, the test sample set was imported into the trained model to test the performance of the model. The test results are shown in Figure 17.

In Figure 17, the abscissa is the number of sample sets, the ordinate is the state of the sample, the circle represents the actual label of the sample, and the asterisk represents the prediction label of the sample. N stands for the healthy bearing, IR for the inner race fault, C for the cage fault, and OR for the outer race fault. In this experiment, the fault classification accuracy reaches 95.5556%. The fault classification of the healthy bearing reaches 100%. And the fault diagnosis of the three kinds of fault bearings mainly occurs in the diagnosis of the cage. For example, the fault of the inner race or the outer race is misdiagnosed as the fault of the cage, or the fault of the cage is misdiagnosed as the fault of the inner race or the fault of the outer race. This indicates that the characteristics of the three faults are similar to some extent. However, on the whole, the method proposed in this paper still has certain advantages in fault diagnosis on the
XJTU-SY bearing data set, which also proves the generalization of the method proposed in this paper.

5. Conclusion and Future Work

This paper takes the bearing fault diagnosis of the gearbox as the research object. Firstly, the vibration signals are extracted by multiscale dispersion entropy (MDE). Secondly, they are imported into the kernel extreme learning machine (KELM) for fault classification. A novel krill herd algorithm (NKH) with opposite-based learning (OBL) and impulse operators is used to optimize kernel function parameters $\sigma$ and the error penalty factor $C$ in the KELM. Then, the bearing data from Case Western Reserve University are imported into NKH-KELM for training and testing. Finally, the method proposed in this paper is compared with the other seven methods. It can be seen from the experimental results that NKH-KELM is superior to the other seven methods in the diagnosis of three fault types, which also proves the superiority of NKH-KELM in intelligent fault diagnosis of bearings. By comparing NKH-KELM with standard KH+KELM, it can be seen that the optimization ability of the novel krill herd algorithm is better than that of the unimproved krill herd algorithm. However, in the diagnosis of ball faults, although NKH-KELM is better than the other seven algorithms, the diagnostic accuracy of ball faults is significantly lower than that of the other two fault types. This phenomenon may be related to the low signal-to-noise ratio in the acquisition of the vibration signal of the ball, the small proportion of useful information on the ball extracted by the feature extraction method (MDE), or the poor performance of machine learning in the fault diagnosis of the ball. Therefore, this question will be an important research direction in the future. In addition, in the era of big data, deep learning methods are becoming more and more perfect, so not only is there a great demand for intelligent fault diagnosis based on machine learning but also there will be a great demand for intelligent fault diagnosis based on deep learning, which will also be a key research direction.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outer race diameter</td>
<td>39.80 mm</td>
<td>Inner race diameter</td>
<td>29.30 mm</td>
</tr>
<tr>
<td>Bearing mean diameter</td>
<td>34.55 mm</td>
<td>Ball diameter</td>
<td>7.92 mm</td>
</tr>
<tr>
<td>Number of balls</td>
<td>8</td>
<td>Contact angle</td>
<td>0°</td>
</tr>
<tr>
<td>Load rating (static)</td>
<td>6.65 kN</td>
<td>Load rating (dynamic)</td>
<td>12.82 kN</td>
</tr>
</tbody>
</table>

Figure 16: XJTU-SY fault diagnosis laboratory.

Figure 17: Fault diagnosis based on the XJTU-SY bearing data set: results of NKH-KELM with classification accuracy = 95.5556%.

Table 5: Parameters of testing bearing.
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 they have no conflicts of interest.

Acknowledgments
This work was supported in part by the National Natural Science Foundation of China under Grant no. 51905496, in part by the Shanxi Provincial Natural Science Foundation of China under Grant no. 51905496, in part by the Science Foundation of the North University of China under Grant no. XJJ201802.

References


[58] C. Yin, S. Dadras, X. Huang, Y. Cheng, and H. Malek, "The design and performance analysis of multivariate fractional-


