Hindawi
Shock and Vibration
Volume 2021, Article ID 6675078, 11 pages
https://doi.org/10.1155/2021/6675078

Research Article

A New Support Vector Regression Model for Equipment Health Diagnosis with Small Sample Data Missing and Its Application

Qinming Liu,[1] **Wenyi Liu**,[1] **Jiajian Mei**,[1] **Guojin Si**,[2] **Tangbin Xia**,[2] **and Jiarui Quan**[1]

[1]*Department of Industrial Engineering, Business School, University of Shanghai for Science and Technology, 516 Jungong Road, Shanghai 200093, China*
[2]*State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, SJTU-Fraunhofer Center, Shanghai 200240, China*

Correspondence should be addressed to Qinming Liu; lqm0531@163.com

Received 4 December 2020; Revised 7 January 2021; Accepted 8 February 2021; Published 25 February 2021

Academic Editor: Gerardo Silva-Navarro

Actually, it is difficult to obtain a large number of sample data due to equipment failure, and small sample data may also be missing. This paper proposes a novel small sample data missing filling method based on support vector regression (SVR) and genetic algorithm (GA) to improve equipment health diagnosis effect. First, the genetic algorithm is used to optimize support vector regression, and a new method GA-SVR can be proposed. The GA-SVR model is trained by using other data of the variable to which the missing data belongs, and the single-variable prediction method can be obtained. The correlation analysis is used to reconstruct the training set, and the GA-SVR is trained by using the data of the variables related to the missing data to obtain the multivariate prediction method. Then, the dynamic weight is presented to combine the single-variable prediction method with the multiple-variable prediction method based on certain principles, and the missing data are filled with the combined prediction methods. The filled data are used as input of GA-SVM to diagnose equipment failure. Finally, a case study is given to verify the applicability and effectiveness of the proposed method.

1. Introduction

For equipment health diagnosis, complete monitoring data is the premise and foundation for an accurate diagnosis. However, in the actual engineering application, many monitoring sample data are incomplete, including small sample, unbalanced sample, and sample data missing. In the collection of sample data, equipment may not be able to operate normally due to fault, or it can be affected by the environment, and the effective monitoring data collected is less, resulting in less failure sample data. The sample data may also be missing due to abnormal data transmission, sensor repair and replacement, or human factors. This paper importantly considers the condition of small sample data missing.

Recently, with the rapid development of technology, equipment health diagnosis has been widely concerned by a large number of experts and scholars. The intelligent diagnosis methods applied to equipment health diagnosis mainly include expert system (ES), neural networks (NNs), and support vector machine (SVM).

For the expert system, Husain [1] expanded the fault diagnosis of the power transformer, proposed a fuzzy logic expert system for early fault diagnosis of the transformer, and improved the shortcomings of traditional transformer fault diagnosis methods. Berredjem and Benidir [2] proposed a fuzzy expert system based on an improved range overlap method and similarity division method to solve the problem of high noise in bearing fault data. The system was used to realize accurate bearing fault diagnosis, and the feasibility of the model was verified by an example analysis. Cheriet et al. [3] proposed an expert system based on fuzzy logic, which used stator current signal pair for fault diagnosis, and verified the feasibility of the expert system for fault diagnosis of doubly fed wind turbines through simulation experiments. Xu et al. [4] carried out a series of

researches on the fault diagnosis of marine diesel engines, proposed a diagnosis expert system based on belief rules, and applied the proposed method to the abnormal wear detection of marine diesel engines, indicating that the method had good accuracy and stability. Equipment health diagnosis method based on the expert system can acquire knowledge from diagnosis examples, but this method does not have the ability to automatically acquire new knowledge, and the fault tolerance is relatively poor. Thus, the fault diagnosis method based on the expert system has great limitations in practical application.

For neural networks, Xing et al. [5] constructed an automatic fault diagnosis method for reciprocating compressors based on information entropy and radial basis function neural networks. The test results showed that the fault diagnosis method can effectively improve the accuracy of automatic fault diagnosis and the practicability of the condition monitoring system. Yang et al. [6] analyzed the fault diagnosis of rotating machinery, proposed an intelligent diagnosis method based on long-term and short-term memory recurrent neural network, and detected and classified the fault with the help of the correlation between time and space. Gunerkar et al. [7] established a rolling bearing fault diagnosis model based on an artificial neural network (ANN) and applied wavelet transform to preprocess the original signal to extract fault features. ANN and the k-nearest neighbor were used for fault classification of rolling bearing, and the validity of the model was verified by test. In order to solve the problem of end-to-end fault diagnosis of rotating machinery, Wu et al. [8] constructed a one-dimensional CNN model which can directly learn features from the original signal, applied it to the fault diagnosis of the fixed gearbox and planetary gearbox, and showed that the model had high diagnostic accuracy. Han et al. [9] proposed a method for fault diagnosis of the planetary gearbox by using an expanded neural network, which expanded the receiving domain by two times, so as to enhance the learning ability of fault features and improve the diagnosis accuracy. The fault diagnosis method based on an artificial neural network often needs a large number of fault samples to train the neural network, but it is difficult to obtain enough fault data in practical engineering applications. In addition, the neural network has the disadvantages of slow convergence, overfitting, and ease to fall into the local optimal value, which will have a negative impact on the diagnostic accuracy of the equipment.

For the support vector machine, Huang and Fei et al. [10, 11] used the SVM model for equipment fault diagnosis and verified that the model has high accuracy and good generalization ability. Yang et al. [12] established an SVM fault classification model using an ant colony algorithm and verified the effectiveness of the model. Zhang et al. [13] combined SVM with an improved imperialist competitive algorithm and applied it to fault diagnosis of the oil-immersed transformer. The results showed that the method was feasible and effective. Yan and Jia [14] proposed a fault recognition algorithm based on optimized multidomain feature SVM. The feature vectors of fault samples were extracted from the time domain, frequency domain, and time-frequency domain. And Laplace fractional algorithm was introduced to filter fault features. Zhong et al. [15] established a diagnosis model based on convolutional neural network transmission learning and SVM and verified the effectiveness of the model through an example. For the accuracy of transformer fault diagnosis, Huang et al. [16] proposed a diagnosis method based on an improved gray wolf algorithm and SVM. The differential evolution mechanism was introduced into the gray wolf optimization algorithm to improve its performance, and then the SVM optimized by the improved gray wolf algorithm was used for fault diagnosis of the transformer.

Equipment fault diagnosis under the condition of incomplete data also has certain research and development. Zhang and Dong [17] proposed an online nonimputation reasoning method based on mixed Gaussian output for fault detection and identification and proved that the method can accurately identify the fault. Mao et al. [18] studied the bearing fault diagnosis with unbalanced data and constructed an online fault prediction method based on an extreme learning machine. The simulation experiment showed that the method can obtain high fault diagnosis accuracy. Liu et al. [19] proposed a Bayesian network parameter learning method based on BPNN and maximum likelihood estimation to solve the problem of solar-assisted heat pump fault diagnosis under the condition of lack of small sample data and lack of expert knowledge. BP neural network was used to predict and fill in the missing sample data, and the effectiveness of the method was verified by simulation. Chen et al. [20] constructed a fault diagnosis model of missing data based on transfer learning for the fault diagnosis problem with too small complete sample size, an appropriate migration learning mechanism was established to improve the accuracy of fault diagnosis, and the effectiveness of this method was verified by data. Zhao et al. [21] constructed a rolling bearing fault diagnosis model based on normalized CNN under unbalanced data and eliminated the difference of feature distribution by batch normalization. The experimental results showed that the model has a good diagnosis effect and robustness for rolling bearing fault diagnosis under unbalanced data. Qian and Li [22] established a kind of unbalance robust network for bearing fault diagnosis, which was used to solve the class imbalance problem in the feature extraction stage and classification stage, and the method was verified by simulation analysis. Zhang et al. [23] proposed to use the deep learning method to solve the problem of fault diagnosis when the data was unbalanced and established a deep generated countermeasure network to generate false samples to balance the sample data. Simulation experiments showed that the proposed method has a better effect on fault diagnosis under unbalanced data.

Collecting sample data in the field of fault diagnosis, a large number of fault sample data cannot be obtained because equipment may not operate normally due to the existence of faults. Presently, most of the research on equipment fault diagnosis is based on complete data set, the research on equipment fault diagnosis under incomplete data is less, and there are some problems such as complex

diagnosis process, long diagnosis time, and unsatisfactory accuracy.

Small sample data missing can not only increase the difficulty of data analysis but also greatly affect the accuracy of the equipment failure diagnosis. For most of equipment failure diagnosis under data missing, it needs a large number of failure sample data to obtain more accurate diagnosis results. Actually, due to equipment aging or human error, a large number of sample data cannot be collected, and there is sample data missing. Thus, the objective of this paper is to propose a novel small sample data missing filling method based on GA-SVR to improve the equipment failure diagnosis effect.

For equipment fault diagnosis, ANN needs a large number of failure samples to train the neural network, but it is difficult to obtain enough failure data in practical application. Additionally, the neural network has the disadvantages of slow convergence, overfitting, and ease to fall into the local optimal value. These will have an adverse impact on the diagnostic accuracy of equipment. Actually, equipment may not operate normally due to failure. And it is unable to obtain a large number of failure sample data. SVR needs less training samples and has high model accuracy. Thus, it is suitable for equipment fault diagnosis in the case of small samples. The advantages of GA lie in its fast optimization speed, good effect, and strong global search ability, and it is not easy to fall into the local optimal solution. Thus, it is used to optimize the key parameters of SVR. In this paper, first, the GA-SVR model is trained by using other data of the variable to which the missing data belongs, and the single-variable prediction method can be obtained. The correlation analysis is used to reconstruct the training set, and the GA-SVR is trained by using the data of the variables related to the missing data to obtain the multivariate prediction method. Then, the dynamic weight is presented to combine the single-variable prediction method with the multiple-variable prediction method based on certain principles, and the missing data are filled with the combined prediction methods. The filled data are used as input of GA-SVM to diagnose equipment failure. Finally, a case study is given to verify the applicability and effectiveness of the proposed method.

This paper aims to develop a new method for equipment health diagnosis. The paper is organized as follows. In section 2, the basic theories of SVR and GA are introduced. Section 3 develops a novel GA-SVR. In Section 4, a case study for equipment health diagnosis with small sample data missing is analyzed and discussed. Finally, conclusions are drawn in Section 5.

2. Theoretical Background

2.1. Support Vector Regression. For the support vector regression (SVR), it is to use the given sample data to fit a continuous function which can reflect the relationship between input and output. In the case that the sample is linear and inseparable, SVR uses a nonlinear transformation to map the data set to a high-dimensional space and carries out regression fitting in this space to establish the continuous function with the minimum loss function.

The key parameters of SVR include insensitive loss function ε, radial basis function parameter σ, and penalty factor C. ε represents the insensitive region width and plays a decisive role in the number of support vectors and the generalization ability of the model. σ determines the complexity of sample mapping space. The larger σ means that it is difficult to obtain high regression accuracy. The smaller σ means that the regression accuracy is high and the generalization ability is poor. C represents the penalty degree for samples with an error greater than ε. The larger C indicates that the penalty for samples is large. Although the training accuracy can be improved, the generalization ability of the model is poor. The smaller C shows that the penalty for samples is very small, and it will cause a large training error. These three key parameters determine the performance of SVR; thus, it is necessary to optimize these parameters to improve the prediction effect of SVR.

2.2. Genetic Algorithm. Genetic algorithm (GA) is a kind of heuristic optimization technology. GA searches from the initial population generated randomly, and the individuals in the population evolve through selection, crossover, and mutation based on the fitness function until the iteration termination condition is met, and the optimal solution is output.

The advantages of GA include fast optimization speed and strong global searchability, and it is not easy to fall into the local optimal solution. It is widely used in various optimization problems such as parameter optimization and path optimization.

The basic procedure of GA is as follows:

Step 1. The chromosome needs to be coded to determine the initial population

Step 2. The fitness function is described to evaluate the fitness value of individuals

Step 3. The new species group is generated by selection, crossover, and mutation

Step 4. The individuals satisfied the termination iteration condition that can be retained

Step 5. The decoding outputs the global optimal solution

In this paper, for the problem of equipment health diagnosis, SVR is used to predict and fill the missing data. But the values of kernel function parameter σ, penalty factor C, and insensitive loss function ε in SVR are particularly important. Thus, the set of key parameters (C, σ, ε) of SVR can be regarded as a population, and the key parameters of SVR can be optimized by GA to improve the prediction performance of SVR.

3. Equipment Health Diagnosis Based on GA-SVR

3.1. Support Vector Regression Optimized by Genetic Algorithm. SVR is obtained by introducing insensitive loss function into SVM. It is usually used to solve regression

fitting problems and seek a regression function representing the relationship between input and output.

For the given data set $\{x_i, y_i\}$, $i = 1, 2, \ldots, N$, where $x_i \in R^n$ is the input sample, and $y_i \in R$ is the output expected value. Assume that SVR maps samples to a high-dimensional space by nonlinear transformation $\phi(*)$ to establish the regression function, and it is as follows:

$$f(x) = w \cdot \phi(x) + b, \tag{1}$$

where w and b are regression function coefficients. And insensitive loss function ε is introduced and defined as

$$L_\varepsilon(f(x), y) = \begin{cases} |y - f(x)| - \varepsilon, & |y - f(x)| \geq \varepsilon \\ 0, \text{ other} \end{cases} \tag{2}$$

Thus, the objective function can be defined as $\min (1/2)\|w\|^2$, and the constraints are

$$\begin{cases} y_i - w \cdot x_i - b \leq \varepsilon, \\ w \cdot x_i + b - y_i \leq \varepsilon, \end{cases} \quad i = 1, \ldots, N. \tag{3}$$

The relaxation factors ξ_i and ξ_i^* are introduced under the condition of allowing the fitting error; then, the objective function is

$$\min\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)\right),$$

$$\text{S.T.} \begin{cases} y_i - w \cdot x_i - b \leq \varepsilon + \xi_i, \\ w \cdot x_i + b - y_i \leq \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, \end{cases} \tag{4}$$

where $C > 0$ is the penalty factor, and it is used to control the punishment for errors exceeding ε. By introducing the Lagrange multiplier α_i and α_i^*, then the above problem is transformed into its dual problem.

$$\max\left\{\sum_{i=1}^{N}(\alpha_i^* - \alpha_i)y_i - \sum_{i=1}^{N}(\alpha_i^* + \alpha_i)\varepsilon - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^* - \alpha_i)\left(\alpha_j^* - \alpha_j\right)K\left(x_i, x_j\right)\right\},$$

$$\text{S.T.} \left\{\sum_{i=1}^{N}\left(\alpha_i^* - \alpha_i\right) = 0, 0 \leq \alpha_i \leq C, 0 \leq \alpha_i \leq C, i = 1, 2, \ldots, N, \right. \tag{5}$$

where $K(x_i, x_j) = \phi(x_i)\phi(x_j)$ is the kernel function. By solving equation (5), the regression fitting function can be obtained as follows:

$$f(x) = \sum_{i=1}^{N}\left(\alpha_i - \alpha_i^*\right)K\left(x_i, x_j\right) + b. \tag{6}$$

For the selection of the SVR kernel function, the RBF kernel function is used in this paper, and its parameter $\sigma > 0$ is the kernel function width factor. It has an important influence on the regression prediction effect of SVR.

The small sample data missing has a great influence on the equipment diagnosis results; thus, this paper uses SVR to execute regression fitting for the missing data. However, the key parameters C, σ, and ε have a great influence on the regression prediction accuracy of SVR. GA is used to optimize C, σ, and ε to improve the prediction performance of SVR for missing data.

The optimization process of C, σ, and ε by GA can be shown in Figure 1, and the specific operation steps are as follows:

Step 1. Parameter initialization: initialize GA parameters and C, σ, and ε; any group (C, σ, ε) represents an individual in GA.

Step 2. Fitness value calculation: in order to evaluate the advantages and disadvantages of GA in selecting SVR parameters, the K-fold cross-validation method is used to take the mean value of K-th root mean square error as the fitness value of an individual, and the calculation of fitness value is as follows:

$$F = \frac{1}{K}\sum_{i=1}^{K}\sqrt{\frac{\sum_{j=1}^{n}(y - \hat{y})^2}{n}}. \tag{7}$$

Step 3. Terminating iteration: if the condition of terminating iteration has not been reached, the selection, crossover, and mutation will be carried out to generate a new group; then, go back to Step 2 to continue iteration.

Step 4. Output optimal values: the optimal values of C, σ, and ε are output after completing iteration and obtain the GA-SVR model.

3.2. Combination Prediction Filling Based on GA-SVR

3.2.1. Single-Variable Prediction Filling Based on GA-SVR. The monitoring data of equipment operation status is mostly time series. It is a series of monitoring values X_t^q obtained by multiple sensors in a time sequence where $t\,(t = 1, 2, \ldots, n)$ represents t-th time point, $q\,(q = 1, 2, \ldots, m)$ denotes the q-th sensor, and X_t^q means the monitoring data value corresponding to the q-th sensor at the t-th time point.

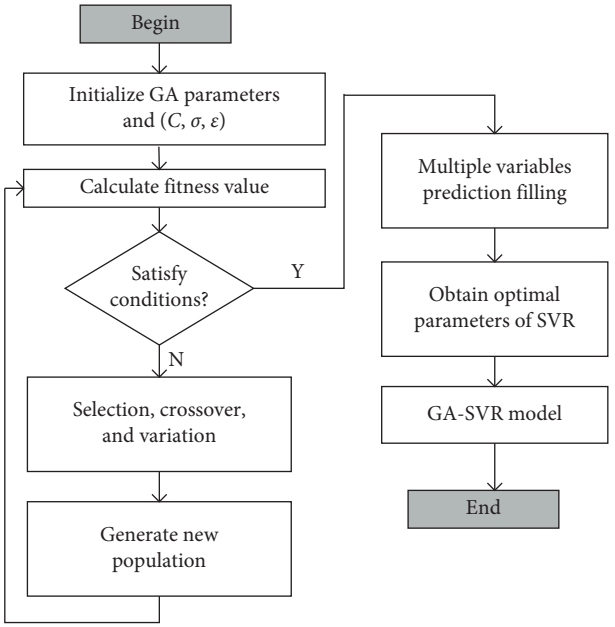

Figure 1: The flow chart of SVR parameters optimized by GA.

Using GA-SVR to predict the single variable of missing data is to train GA-SVR by using other data of variables with missing data as input to predict the value of missing data.

First, let the length of the missing data segment be l, and determine the variable q of the missing data. The n-l-1 data values in the q-th variable dimension are selected as the input of GA-SVR, and the remaining data value is used as the output to train GA-SVR. Then, the trained GA-SVR model is used to predict missing data, and the single-variable prediction results can be obtained.

3.2.2. Multiple-Variable Prediction Filling Based on GA-SVR. This paper uses GA-SVR to predict the missing data. The data related to the variable dimension containing missing data is used as input to train the GA-SVR model and predict the value of missing data.

First, the correlation analysis is used to find the other variables related to the variable q to form the training set $X_1, \cdots, X_i, \cdots, X_k$. X_t represents the monitoring value at t-th time point. The correlation coefficient R is used to evaluate the correlation among the variables. If the correlation coefficient $R \geq 0.8$, it indicates that the two variables are strongly correlated. The correlation coefficient R is calculated as follows:

$$R = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}. \tag{8}$$

The monitoring data from 1-st to k-th time point can be used to execute correlation analysis. And the GA-SVR is trained with the monitoring data values at remaining n-k time points as the input and the data values at a time point where the missing data belongs to as the output. Then, the trained GA-SVR model is used to predict the missing data and obtain the multivariable prediction results.

3.2.3. Dynamic Weight Combination Prediction Filling Based on GA-SVR. In order to improve the accuracy of missing data prediction, reduce the deviation between the predicted value and the actual value, and improve the effectiveness of equipment fault diagnosis, a dynamic weight combination prediction method based on GA-SVR is established to fill the missing data. GA-SVR is used to make a single-variable prediction and multiple-variable predictions, respectively, and then the dynamic weight combination of single-variable prediction and multiple-variable prediction results is obtained. The combined prediction results are used to fill in the missing data to obtain complete data set.

Root mean square error (RMSE) can describe the deviation between the predicted value and the actual value. Thus, RMSE is used to evaluate the quality of the prediction results. The smaller RMSE represents the better prediction effect of missing data. The root mean square error is expressed as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{n} (y_i - \widehat{y}_i)^2}{n}}, \tag{9}$$

where y_i is the actual value, $\widehat{y}_i$ is the predicted value, and n is the prediction times.

The weight value of single-variable prediction results and multiple-variable prediction results in combination forecasting depends on their root mean square error difference. The root mean square error is smaller, and the weight is greater. Based on equation (10), the prediction result of missing data can be obtained and it is followed as equation (11).

$$\begin{cases} w_1 = \dfrac{k}{R_1}, \\ w_2 = \dfrac{k}{R_2}, \\ w_1 + w_2 = 1, \\ y_i^* = w_1 \widehat{y}_{1i} + w_2 \widehat{y}_{2i}, \end{cases} \tag{10}$$

$$y_i^* = \frac{R_2}{R_1 + R_2} \widehat{y}_{1i} + \frac{R_1}{R_1 + R_2} \widehat{y}_{2i}, \tag{11}$$

where $\widehat{y}_{1i}$ and $\widehat{y}_{2i}$ denote single-variable prediction results and multiple-variable prediction results, respectively. R_1 and R_2 are the RMSE values corresponding to single-variable prediction and multiple-variable prediction, respectively. y_i^* is the final missing data filling values.

The chart of combination prediction based on GA-SVR can be seen in Figure 2.

3.3. Equipment Failure Diagnosis Procedure. For the problem of equipment fault diagnosis under the condition of small sample data missing, GA-SVR is used to fill the missing data, and the complete data after filling is used as the input of SVM to realize the fault diagnosis of equipment. The

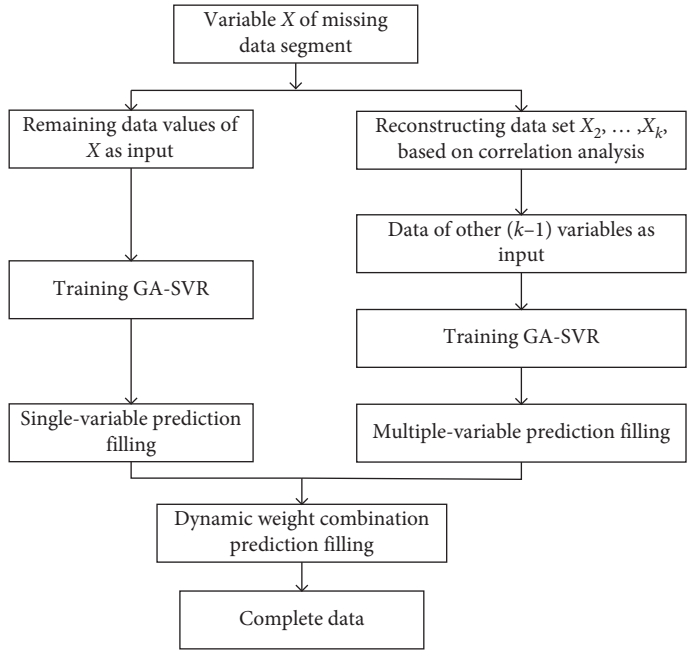

Figure 2: The flow chart of combination prediction based on GA-SVR.

fault diagnosis flow chart based on SVR under the condition of small sample data missing can be shown in Figure 3, and the specific fault diagnosis scheme can be shown as follows:

Step 1. The other data of the variable to which the missing data belongs is used to train the GA-SVR model to obtain the single-variable prediction filling result that can be obtained.

Step 2. Find out the variables related to the variables of missing data by correlation analysis, and the data of these variables can be used to train the GA-SVR model. The multiple-variable prediction filling results can be obtained.

Step 3. Based on equation (11), the single-variable prediction results and the multivariate prediction results are combined to obtain the combined prediction results, and the missing data are filled to obtain the complete data.

Step 4. The complete data is divided into training sample data set and test sample data set, and SVM is trained and tested, respectively, to obtain the fault diagnosis results of equipment.

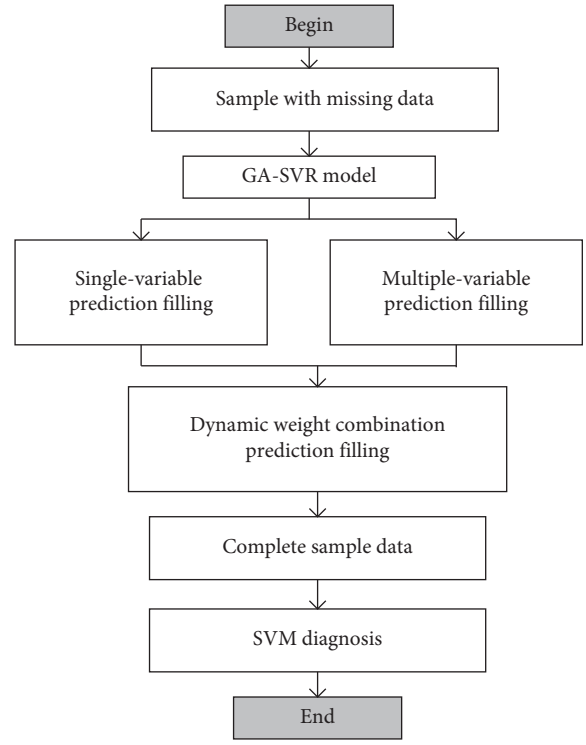

Figure 3: Equipment fault diagnosis scheme based on GA-SVR with small sample data missing.

4. Case Study

4.1. Experimental Setup and Data Acquisition. To validate the proposed methods, a real-world case is studied. In this case study, the long-term wear test experiments were conducted at a research laboratory facility. In the test experiments, three pumps (A, B, and C) were worn by running them using oil containing dust. Each pump experienced four states: Baseline state, Degradation state, Degradation state, and Failure state. The degradation stages in this hydraulic pump wear test case study correspond to different stages of flow loss in the pumps. As the flow rate of a pump clearly indicates the pump's health state, the degradation stages corresponding to different degrees of flow loss in a pump were defined as the health states of the pump in the test [24, 25].

The vibration signals were collected from pump accelerometers that were positioned parallel to the axis of the

swash plate swivel axis and data was continuously sampled. Figure 4 shows the schematic diagram of the experimental setup. The pump used for testing in the experiments was a Back Hoe Loader: a 74 cc/rev variable displacement pump. The data was collected at a sample rate of 60 kHz with antialiasing filters from accelerometers designed to have a usable range of 10 kHz. In many cases, the most distinguished information is hidden in the frequency content of signals. So, the time-frequency representation of signals is needed. In this case study, the signals were processed using a wavelet packet with Daubechies wavelet 10 (db10), and five decomposition levels as the db10 wavelet provide the most effective way to capture the fault information in the pump vibration data. The coefficients obtained by the wavelet packet decomposition were used as the inputs.

There are 80 groups of experimental data for Pumps A, B, and C, respectively. Each group of data contains 32 variables (32 sensors). In this paper, the monitoring data of the 3-th sensor is taken as the experimental object, and the monitoring data from the 75-th to 80-th time point is deleted to simulate the missing situation of small sample data. The single-variable prediction, multiple-variable prediction, and dynamic weight combination prediction based on GA-SVR are used to fill the missing data, and the filling effect and the diagnosis effect after filling are compared.

4.2. Reconstruction Training Set. The multiple-variable prediction model selects monitoring data from sensors having a strong correlation with Sensor 3 as the training set to predict the missing data value. Based on equation (8), the correlation coefficients between Sensor 3 and other sensors are calculated in Pumps A, B, and C, respectively. If the correlation coefficient $R \geq 0.8$, then the sensor and Sensor 3 have a strong correlation; thus, the training set can be reconstructed, as shown in Tables 1–3. The reconstructed training sample is only 6-dimensional. It can reflect the characteristics of the original data, reduce the amount of calculation, and shorten the prediction time.

4.3. Result Analysis of Missing Data Filling. In order to evaluate the filling effect of the proposed dynamic weight combination prediction method based on GA-SVR, the missing values in Pumps A, B, and C are predicted by single-variable prediction, multiple-variable prediction, and dynamic weight combination prediction by using GA-SVR, respectively. And the filling effects are compared.

The parameters of GA are set as follows: the population size is 20, and the maximum iteration number is 100. The key parameters of SVR are $0.1 \leq C \leq 1000$, $0.01 \leq \sigma \leq 100$, and $0.01 \leq \varepsilon \leq 1$. The root mean square error (RMSE) and mean absolute percentage error (MAPE) are used as the evaluation indexes for the filling effect of missing data. MAPE is as follows:

$$\text{MAPE} = \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times \frac{100\%}{n}, \tag{12}$$

where y_i is the actual value and $\hat{y}_i$ is the predicted value.

Tables 4–6 show the predicted filling values of missing data of Pumps A, B, and C based on GA-SVR, respectively. Figures 5–7 show the missing data fitting curves of three prediction methods based on GA-SVR for Pumps A, B, and C, respectively.

From Figures 5–7, it can be intuitively seen that the simulation results of the three data sets are basically consistent. The fitting curve of dynamic weight combination prediction is more consistent with the actual value curve than that of single-variable prediction and multiple-variable prediction. It indicates that the effect of the dynamic weight combination prediction method is better than that of single-variable prediction and multiple-variable prediction.

In order to evaluate the effect of equipment fault diagnosis under the small sample data missing based on the proposed GA-SVR, the proposed GA-SVR prediction model is compared with the standard SVR prediction model and BP neural network prediction model (BPNN). The key parameters of SVR are selected by grid search cross-validation method, $0.1 \leq C \leq 1000$, $0.01 \leq \sigma \leq 100$, and $0.01 \leq \varepsilon \leq 0\,1$. For the single-variable prediction of missing data, the input layer of BPNN is 1, the output layer is 1, and the number of hidden layers is 3. For the multiple-variable prediction of missing data, the input layer of BPNN is 6, the output layer is 1, and the number of hidden layers is 5. The maximum iteration times are set to 100, the error accuracy is 0.002, the learning rate is 0.1, and the activation function is a sigmoid type function.

Tables 7–9 show the filling effect of missing data of Pumps A, B, and C for three different prediction models, respectively. It can be seen from Tables 7–9 that the RMSE and MAPE values of dynamic weight combination prediction are the smallest compared with single-variable prediction and multiple-variable prediction for different prediction modes of the same prediction model. For the same prediction mode of different prediction models, the RMSE and MAPE values of the proposed GA-SVR model are the minimum. Thus, the proposed dynamic weight combination prediction of missing data based on GA-SVR has the best filling effect on missing data.

4.4. Result Analysis of Equipment Failure Diagnosis. In order to compare the effects of different missing data prediction models and prediction modes on equipment fault diagnosis, the complete data filled with missing data is used for equipment fault diagnosis. 50 groups of Pumps A, B, and C data sets are randomly selected as training samples, and the remaining 30 groups are used as test samples.

Tables 10–12 show the influence of three different missing data filling models of GA-SVR, SVR, and BPNN and three prediction filling modes on the fault diagnosis effect of Pumps A, B, and C, respectively. It can be seen from Tables 10–12 that the dynamic weight combination prediction filling mode has the highest diagnosis accuracy rate and shorter time compared with single-variable prediction filling mode and multiple-variable prediction filling mode under the same prediction model. For the same prediction mode, the fault diagnosis rate based on GA-SVR is the

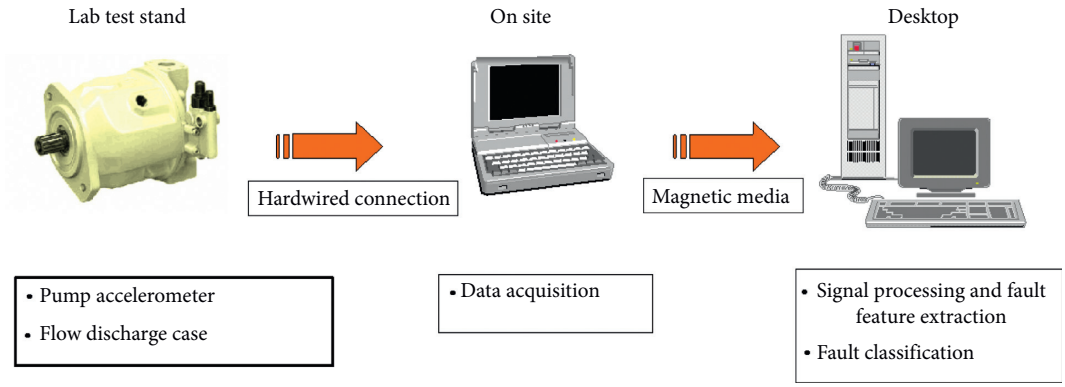

Figure 4: Schematic diagram of the experimental setup.

Table 1: Sensors having a strong correlation with Sensor 3 in hydraulic Pump A.

	CH2	CH5	CH7	CH13	CH16	CH32
R	0.826	0.872	0.908	0.911	0.956	0.858

Table 2: Sensors having a strong correlation with Sensor 3 in hydraulic Pump B.

	CH2	CH5	CH7	CH13	CH16	CH32
R	0.819	0.863	0.895	0.902	0.938	0.838

Table 3: Sensors having a strong correlation with Sensor 3 in hydraulic Pump C.

	CH2	CH5	CH7	CH13	CH16	CH32
R	0.821	0.867	0.901	0.907	0.944	0.843

Table 4: Prediction results of missing data based on GA-SVR for Pump A.

Actual value	Single-variable predicted value	Multiple-variable predicted value	Dynamic weight combination predicted value
16.9640	16.9023	17.0052	16.9502
16.8942	16.8425	16.9732	16.9033
16.7349	16.7745	16.6997	16.7397
16.6608	16.7177	16.6369	16.6801
16.6291	16.6791	16.6002	16.6424
16.7138	16.7330	16.7265	16.7300

Table 5: Prediction results of missing data based on GA-SVR for Pump B.

Actual value	Single-variable predicted value	Multiple-variable predicted value	Dynamic weight combination predicted value
15.1519	15.3855	14.9987	15.2222
14.2496	13.7533	14.4974	14.0675
12.8249	12.5942	13.0492	12.7863
12.9940	12.5854	13.1238	12.8128
12.3819	12.5935	11.9923	12.3396
12.4991	12.8678	12.2324	12.5995

Table 6: Prediction results of missing data based on GA-SVR for Pump C.

Actual value	Single-variable predicted value	Multiple-variable predicted value	Dynamic weight combination predicted value
9.4516	9.3048	9.5537	9.4079
9.3964	9.3058	9.4623	9.3706
9.7349	9.6812	9.7615	9.7145
9.1048	9.1879	9.0531	9.1321
9.5237	9.5981	9.4552	9.5389
9.6634	9.6289	9.6801	9.6501

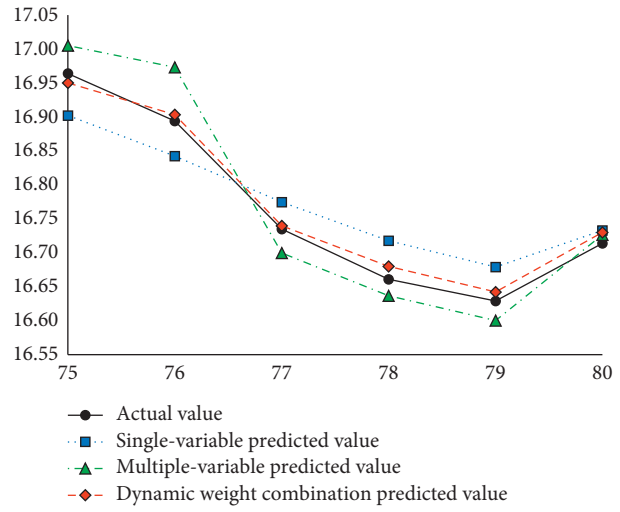

Figure 5: The fitting curve of missing data based on GA-SVR for Pump A.

highest compared with SVR and BPNN, and the diagnosis time is shorter than that of BPNN. And the diagnosis time is longer than SVR, but the difference is not significant.

Generally, the missing data filling method of dynamic weight combination prediction based on GA-SVR can obtain the best fault diagnosis effect. It can be concluded that the proposed failure diagnosis method based on GA-SVR under the condition of small sample missing data is effective for Pumps A, B, and C and has certain universality.

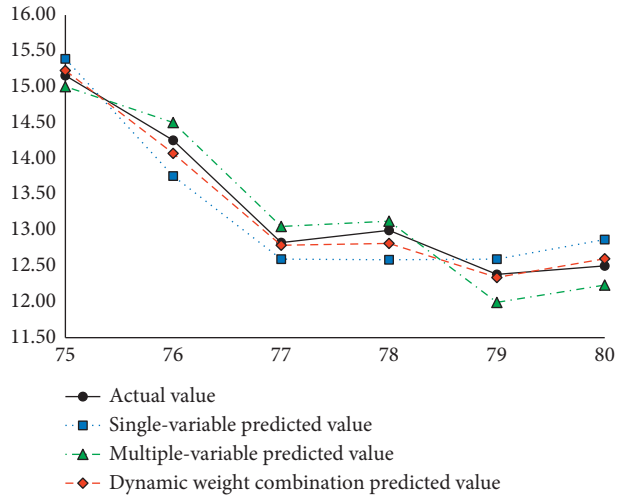

Figure 6: The fitting curve of missing data based on GA-SVR for Pump B.

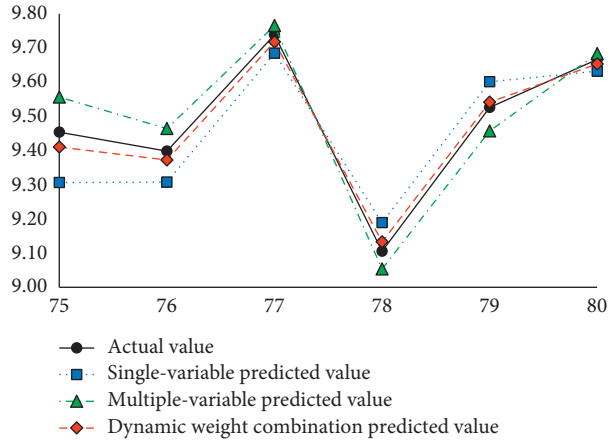

Figure 7: The fitting curve of missing data based on GA-SVR for Pump C.

Table 7: Prediction effect of missing data of Pump A for three different prediction models.

	Single-variable prediction		Multiple-variable prediction		Dynamic weight combination prediction	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
GA-SVR	0.0486	0.28	0.0423	0.22	0.0138	0.08
SVR	0.0737	0.40	0.0500	0.25	0.0303	0.16
BPNN	0.0920	0.52	0.0644	0.36	0.0547	0.28

Table 8: Prediction effect of missing data of Pump B for three different prediction models.

	Single-variable prediction		Multiple-variable prediction		Dynamic weight combination prediction	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
GA-SVR	0.3420	2.44	0.2500	1.80	0.1185	0.76
SVR	0.6547	2.90	0.3989	2.13	0.2158	1.12
BPNN	0.8832	3.28	0.5150	2.82	0.2990	1.98

Table 9: Prediction effect of missing data of Pump C for three different prediction models.

	Single-variable prediction		Multiple-variable prediction		Dynamic weight combination prediction	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE (%)
GA-SVR	0.0878	0.85	0.0621	0.59	0.0263	0.26
SVR	0.1439	1.69	0.1219	1.33	0.0498	0.73
BPNN	0.2293	2.12	0.1580	1.83	0.0724	1.20

Table 10: Failure diagnosis effect of different missing data filling models for Pump A.

	Single-variable prediction filling mode		Multiple-variable prediction filling mode		Dynamic weight combination prediction filling mode	
	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)
GA-SVR	83.33	20.84	90.00	41.20	96.67	41.64
SVR	80.00	21.33	83.33	39.86	90.00	41.50
BPNN	76.67	50.8	90.00	87.23	93.33	88.92

Table 11: Failure diagnosis effect of different missing data filling models for Pump B.

	Single-variable prediction filling mode		Multiple-variable prediction filling mode		Dynamic weight combination prediction filling mode	
	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)
GA-SVR	86.67	16.43	93.33	23.45	100.00	24.29
SVR	83.33	14.50	90.00	22.76	96.67	23.40
BPNN	76.67	31.80	86.67	59.80	93.33	61.02

Table 12: Failure diagnosis effect of different missing data filling models for Pump C.

	Single-variable prediction filling mode		Multiple-variable prediction filling mode		Dynamic weight combination prediction filling mode	
	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)	Accuracy (%)	Time (s)
GA-SVR	86.67	9.43	93.33	13.45	96.67	14.23
SVR	83.33	8.55	86.67	10.98	93.33	11.45
BPNN	80.00	15.78	83.33	21.50	90.00	23.27

5. Conclusion

In this paper, for the problem that small sample data missing will affect the effect of equipment failure diagnosis, a novel missing data filling method based on GA-SVR is proposed to improve the effect of the equipment failure diagnosis. First, the single-variable prediction is carried out for the missing data. And the training set is reconstructed by correlation analysis. Meanwhile, the multiple-variable prediction is carried out based on GA-SVR. Then, the dynamic weight is presented to combine the single-variable prediction results and the multiple-variable prediction results to fill in the missing data. Finally, the complete data obtained by filling missing data is used as input, and GA-SVM is used to diagnose the equipment failure.

By the case study, the proposed GA-SVR model is compared with SVR and BPNN to predict the filling effect of missing data of Pumps A, B, and C, respectively. And the failure diagnosis effect based on the complete data after the filling is compared. It can be shown that the proposed dynamic weight combination prediction method based on GA-SVR has the best missing data filling effect and failure diagnosis effect. And the effectiveness and universality of this proposed method under the condition of small sample data missing can be verified.

Data Availability

The underlying data supporting the results of our study can be obtained, including, where applicable, hyperlinks to publicly archived datasets analyzed or generated during the study, upon request to the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (Nos. 71632008, 71840003, and 51875359), Natural Science Foundation of Shanghai (Nos. 19ZR1435600 and 20ZR1428600), and

Humanity and Social Science Planning Foundation of the Ministry of Education of China (No. 20YJAZH068).

References

[1] Z. Husain, "Fuzzy logic expert system for incipient fault diagnosis of power transformers," *International Journal on Electrical Engineering and Informatics*, vol. 10, no. 2, pp. 300–317, 2018.

[2] T. Berredjem and M. Benidir, "Bearing faults diagnosis using fuzzy expert system relying on an Improved Range Overlaps and Similarity method," *Expert Systems with Applications*, vol. 108, pp. 134–142, 2018.

[3] A. Cheriet, A. Bekri, A. Hazzab, and H. Gouabi, "Expert system based on fuzzy logic: application on faults detection and diagnosis of DFIG," *International Journal of Power Electronics and Drive Systems (IJPEDS)*, vol. 9, no. 3, pp. 1081–1089, 2018.

[4] X. Xu, X. Yan, C. Sheng et al., "A belief rule based expert system for fault diagnosis of marine diesel engines," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 2, pp. 656–672, 2020.

[5] C. H. Xing, F. T. Xu, Z. Y. Yao et al., "A fault diagnosis method of reciprocating compressor based on sensitive feature evaluation and artificial neural network," *High Technology Letters*, vol. 21, no. 4, pp. 422–428, 2015.

[6] R. Yang, M. Huang, Q. Lu, and M. Zhong, "Rotating machinery fault diagnosis using long-short-term memory recurrent neural network," *IFAC-PapersOnLine*, vol. 51, no. 24, pp. 228–232, 2018.

[7] R. S. Gunerkar, A. K. Jalan, and S. U. Belgamwar, "Fault diagnosis of rolling element bearing based on artificial neural network," *Journal of Mechanical Science and Technology*, vol. 33, no. 2, pp. 505–511, 2019.

[8] C. Wu, P. Jiang, C. Ding, F. Feng, and T. Chen, "Intelligent fault diagnosis of rotating machinery based on one-dimensional convolutional neural network," *Computers in Industry*, vol. 108, pp. 53–61, 2019.

[9] Y. Han, B. Tang, and L. Deng, "An enhanced convolutional neural network with enlarged receptive fields for fault diagnosis of planetary gearboxes," *Computers in Industry*, vol. 107, pp. 50–58, 2019.

[10] J. Huang and X. G. Hu, "Support vector machine with genetic algorithm for machinery fault diagnosis of high voltage circuit breaker," *Measurement*, vol. 44, no. 6, pp. 1018–1027, 2011.

[11] S. W. Fei and X. B. Zhang, "Fault diagnosis of power transformer based on support vector machine with genetic algorithm," *Expert Systems with Applications*, vol. 36, no. 8, pp. 11352–11357, 2009.

[12] D. Yang, Y. Liu, S. Li, X. Li, and L. Ma, "Gear fault diagnosis based on support vector machine optimized by artificial bee colony algorithm," *Mechanism and Machine Theory*, vol. 90, no. 90, pp. 219–229, 2015.

[13] Y. Zhang, H. Wei, R. Liao, Y. Wang, L. Yang, and C. Yan, "A new support vector machine model based on improved imperialist competitive algorithm for fault diagnosis of oil-immersed transformers," *Journal of Electrical Engineering and Technology*, vol. 12, no. 2, pp. 830–839, 2017.

[14] X. Yan and M. Jia, "A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing," *Neurocomputing*, vol. 313, pp. 47–64, 2018.

[15] S.-s. Zhong, S. Fu, and L. Lin, "A novel gas turbine fault diagnosis method based on transfer learning with CNN," *Measurement*, vol. 137, pp. 435–453, 2019.

[16] X. Huang, X. Huang, B. Wang, and Z. Xie, "Fault diagnosis of transformer based on modified grey wolf optimization algorithm and support vector machine," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 15, no. 3, pp. 409–417, 2020.

[17] Z. Zhang and F. Dong, "Fault detection and diagnosis for missing data systems with a three time-slice dynamic Bayesian network approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 138, pp. 30–40, 2014.

[18] W. Mao, L. He, Y. Yan, and J. Wang, "Online sequential prediction of bearings imbalanced fault diagnosis by extreme learning machine," *Mechanical Systems and Signal Processing*, vol. 83, pp. 450–473, 2017.

[19] Z. Liu, Y. Liu, D. Zhang, B. Cai, and C. Zheng, "Fault diagnosis for a solar assisted heat pump system under incomplete data and expert knowledge," *Energy*, vol. 87, pp. 41–48, 2015.

[20] D. Chen, S. Yang, and F. Zhou, "Transfer learning based fault diagnosis with missing data due to multi-rate sampling," *Sensors*, vol. 19, no. 8, p. 1826, 2019.

[21] B. Zhao, X. Zhang, H. Li, and Z. Yang, "Intelligent fault diagnosis of rolling bearings based on normalized CNN considering data imbalance and variable working conditions," *Knowledge-Based Systems*, vol. 199, Article ID 105971, 2020.

[22] W. Qian and S. Li, "A novel class imbalance-robust network for bearing fault diagnosis utilizing raw vibration signals," *Measurement*, vol. 156, Article ID 107567, 2020.

[23] W. Zhang, X. Li, X.-D. Jia, H. Ma, Z. Luo, and X. Li, "Machinery fault diagnosis with imbalanced data using deep generative adversarial networks," *Measurement*, vol. 152, Article ID 107377, 2020.

[24] M. Dong and D. He, "A segmental hidden semi-Markov model (HSMM)-based diagnostics and prognostics framework and methodology," *Mechanical Systems and Signal Processing*, vol. 21, no. 5, pp. 2248–2266, 2007.

[25] K. M. Hancock and Q. Zhang, "A hybrid approach to hydraulic vane pump condition monitoring and fault detection," *Transactions of the ASABE*, vol. 49, no. 4, pp. 1203–1211, 2006.