Applied Intelligence (2021) 51:804–818
https://doi.org/10.1007/s10489-020-01830-0

Design of fault diagnosis algorithm for electric fan based on LSSVM and Kd-Tree

Kongzhi Hu[1] · Ming Jiang[1] · Haifeng Zhang[1] · Sheng Cao[1] · Ziyi Guo[1]

Published online: 2 September 2020

Abstract
Currently, the complexity of mechanical equipment is increasing rapidly together with the poor working environment. If a fault occurs, how to find the fault in time becomes a poser. Motivated by this existing problem, based on the analysis of the fault characteristics of electric fans, a fault diagnosis algorithm model based on Least Square Support Vector Machine (LSSVM) and Kd-Tree was proposed. This algorithm was based on the LSSVM optimized by the Cuckoo Search (CS). This paper used the "one-to-many" principle and the sigma threshold method to introduce k-Nearest Neighbor (kNN) which was implemented by Kd-Tree as a secondary classifier to optimize the model. In data preprocessing, the data based on time series was first processed by Empirical Mode Decomposition (EMD) and the energy ratios were calculated, and the the above results were degraded by Principal Component Analysis (PCA) and normalized. On top of that, in case of the uncertain fault types, the Fuzzy C-Means clustering algorithm (FCM) optimized by Particle Swarm Optimization (PSO) was proposed to provide a priori knowledge for the model. In this paper, the algorithm model, FCM and other parts were verified to prove that the performance and generality of the algorithm were better than those of general classification algorithms, and relevant experiments were conducted for different data processing methods to expand the universality of the algorithm.

Keywords Least Square Support Vector Machine · Kd-Tree · Data preprocessing · Fuzzy C-Means clustering algorithm · Model optimization

1 Introduction

Nowadays, modern industrial production plays an indispensable role to the society due to its conveniences and efficiency. However, such complex mechanical production equipment may have unpredictable errors and faults during long-term and high-speed operation [1–3]. Equipment maintenance is both time-consuming and money-costly, and, worse still, fatal errors can even cause injury or death [4–6]. Therefore, it is necessary to implement a real-time diagnosis system to monitor whether an error is being generated or will occur on the mechanical equipment. Fault diagnosis has received widespread attention after the British doctoral scholar R.A.Collacott published "Structural integrity monitoring" [7]. Now it is further developed and applied, gradually integrating k-Nearest Neighbor [8, 9], Support Vector Machine [10, 11] and many other machine learning classification algorithms [12, 13]. The introduction of BackPropagation Neuron Network (BP Neuron Network) introduces new ideas for fault diagnosis [14, 15]. More and more experts and scholars are investing in research. Such as Liang et al. proposed the use of Recursive Neural Networks to predict mechanical life [16], Cartella et al. used hidden semi-Markov model for state judgment [17], and Buzzoni et al. proposed a new blind deconvolution algorithm based on cyclic stationary [18], which is called maximum second-order cyclic stationary blind deconvolution (CYCBD) for processing bearing signals. Many methods proposed by these scholars provide more new ideas for the follow-up research of fault diagnosis, and mang good results are received in corresponding situations.

Traditional fault diagnosis algorithms often use feature extraction schemes in conjunction with ordinary machine learning strategies to diagnose the state of mechanical devices. In addition, some scholars have reduced the correlation between fault characteristics and working conditions in the study, so that they can extract deeper features and achieved good results [19–21]. But as we all know, these

✉ Ming Jiang
mjiang@hit.edu.cn

[1] School of Astronautics, Harbin Institute of Technology, Harbin, 150001, China

algorithms are not necessarily applicable to all situations, so need to design and modify algorithms according to specific application conditions. Electric fan system has many strange features that may affect the performance of older machine learning algorithms [22]. Against this background, it is of great significance to analyze the characteristics of electric fan system and adopt the appropriate scheme. In the tasks of fault diagnosis, there are situations that can result in performance degradation of traditional machine learning strategies [23]. First, when the distribution of fault data is uneven, the performance of some algorithms (such as the decision tree) will be seriously affected [24]. In addition, the prior knowledge of different failures that humans have acquired is limited, and this prior knowledge may affect the work of the algorithm [25]. The imbalance of data is also an important factor that affects the performance of the algorithm [26–28]. There are many ways to deal with it, such as oversampling, undersampling, penalty processing, etc. The method of unbalanced data processing based on Generative Adversarial Network (GAN) proposed by Amin Bemani et al. has brought scholars a new look [26]. It is also extremely important to select the appropriate data processing method for the algorithm [27]. Due to many challenges, this paper modified some traditional algorithms and designed new strategies to improve the performance of the entire algorithm.

This paper aimed to solve malfunction detection and prediction in electric fan system. A fault diagnosis algorithm that combined the advantages of different algorithms was designed to monitor automatically the state of electric fans, and combined with surveillance system, which provided users a simple way to watch the history state of electric fans and alerts in time when a hardware error occurred. As for the structure of the whole system, it consisted of hardware sensors, wireless network transformation, PC monitor software and fault diagnosis algorithm. This paper provided an accurate description of this algorithm.

The method proposed in this paper aimed to be compatible with the fault diagnosis tasks and improve the performance of traditional algorithms. There are many types of mechanical equipment faults, and the fault diagnosis is actually a multi-classification task. This paper combined the advantages of Least Square Support Vector Machine (LSSVM) and k-Neighbor-Nearest (kNN) to design algorithm models. A preprocessing method was designed for data based on time series. Aiming at the lack of prior knowledge of faults, a processing method based on Fuzzy C-Means clustering algorithm (FCM) was designed to enhance the practicality of the algorithm. In the experimental stage, this paper used the performance of GAN and the characteristics of the LSSVM to conduct experiments and discussions on data imbalance and data simplification [26–28].

The chapters of this article are arranged as follows: In the second section, the structure of the proposed algorithm model and the FCM to provide prior knowledge are described. In the third section, the experimental steps of important parts of the paper are described. The experimental results are presented and the different processing methods of the data are discussed in Section 4. Section 5 is the summary and outlook of the work of this paper.

2 Structure of algorithm model

2.1 LSSVM classifier optimized with CS

In the above, a brief analysis was given according to the difficulties of the fault diagnosis tasks. In order to make the algorithm designed in this paper more suitable for the task, this paper added the idea of least squares to the traditional Support Vector Machine (SVM), and improved it to LSSVM.

Least Squares Support Vector Machine (LSSVM) is an improvement and optimization of traditional SVM. It is a kernel function algorithm model that follows the principle of minimizing structural risks. Training samples $\{(x_k, y_k)|k = 1, 2, ..., l\}$, $x_k \in R^n$ is input sample, $y_k \in \{1, -1\}$ is category of the k-th sample, and is sample size. The classification problem expression of LSSVM in the weight space is as follows:

$$\begin{aligned} &\min_{w,b,e} J(w, e) = \tfrac{1}{2} w^T w + \tfrac{1}{2}\gamma \sum_{k=1}^{N} e_k \\ &s.t.\ y_k \left[w^T \varphi(x_k) + b\right] = 1 - e_k,\ k = 1, \ldots, N \end{aligned} \tag{1}$$

Where, $\varphi(x_k) : R^n \to R^m$ represents a non-linear mapping function relationship, $w \in R^m$ represents a weight vector, $e_k \in R$ represents errors variable, b is a constant deviation. And γ represents an adjustable regularization parameter, which controls penalties for error samples in the data [29]. For the problem of formula (1), after constructing the Lagrange function and introducing the kernel function, the solution expression of LSSVM is finally obtained:

$$y(x) = \text{sign}\left[\sum_{k=1}^{N} \alpha_k y_k K(x, x_k) + b\right] \tag{2}$$

The most widely used radial basis function was chosen:

$$K(x_1, x_2) = \exp\left(-\frac{||x_1 - x_2||}{2\sigma^2}\right), \sigma > 0 \tag{3}$$

Where, σ is its width parameter that controls its radial range.

LSSVM saves training time to a large extent, and the data that is updated in real time can update the model at a faster speed. At the same time, in order to ensure the hyper-parameters of LSSVM and to search for the optimal value

more quickly and accurately. this paper also introduced CS to the hyper-parameter optimization process of LSSVM.

Cuckoo Search (CS) is a new meta-heuristic swarm intelligence optimization algorithm researched and proposed by Yang etc. From Cambridge University. Its idea is mainly based on two strategies: the nest parasitism of cuckoos and the Levy Flights mechanism. This method is a search mode with high performance [30]. Therefore, CS has the characteristics of strong optimization ability, high fitting degree and fast convergence speed.

The training time of the LSSVM algorithm based on CS-optimized search could be greatly shortened, and the precision was guaranteed to be within the acceptable range. The optimization of the hyper-parameters would be faster, skipping the local optimum.

2.2 Implementation and optimization of kNN classifier

Because the fault diagnosis algorithm requires higher accuracy of the results, most of the fault diagnosis is multi-classification tasks. LSSVM functions well on the two-classification task. However, when completing the multi-classification tasks, it is inevitable that the accuracy would decrease. If misjudgment occurs (for example, the equipment works normally, and the algorithm determines that it has failed), it will consume a lot of human and financial resources, frequent inspection, repair of sensors and other components can easily cause damage to the components. In order to prevent the errors of the classifier, this project decided to improve the accuracy and precision of the classifier through model fusion. The kNN algorithm and the LSSVM were used to perform model fusion for decision.

K-Nearest Neighbor (kNN) was proposed by COVER et al. in 1968. It is a classification algorithm based on analogy, which has the characteristics of simple model and high robustness [31]. As the linear scanning method of the KNN algorithm takes a long time, it would greatly affect the execution speed. In order to increase the real-time nature of the algorithm, this project intended to use the K-D tree to replace the original linear scanning to implement the kNN algorithm. On the premise that the spatial dimension is much smaller than the number of training data, the K-D tree could greatly save online search time and meet the real-time requirements of fault diagnosis algorithms [32].

2.3 Algorithm fusion of primary and secondary classifiers

When dealing with multiple classification problems, multiple LSSVMs were trained by dividing the data into "one-to-many" and undersampling some data. The main classifier was formed by fusing multiple LSSVMs in a "one-to-many" style. When constructing the main classifier, it was inevitable that data imbalance was encountered. Even though LSSVM is less sensitive to unbalanced data, it is still affected by it. This paper has adopted an undersampling strategy for this problem. Now, it is well known that Generative Adversarial Network (GAN) is more and more applied to serious unbalanced data problems and has achieved great results [26–28]. The idea of GAN was tried for data augmentation in the experimental stage (Section 4.7) to deal with unbalanced data. Since the main classifier is composed of multiple LSSVMs, the efficiency is inevitably reduced. In Section 4.7, this paper attempted to improve the model speed by filtering data. On the basis of the original undersampling scheme, the data closer to the support vector of LSSVM was selected for relevant experiments, and more attention was paid to the data balance in this process.

On the fusion of the K-D Tree and the main classifier, because LSSVMs were fused in a "one-to-many" manner to complete the multi-classification tasks, individual vector machines would make errors, and when training the main LSSVM classifier, large errors might be caused by the points located near the hyperplane, so this project proposed the following two fusion strategies according to the credibility of algorithm:

(1) KNN retest. When the output of more than one LSSVM in the main classifier were a positive class, the K-D tree was used to recheck.
(2) Sigma threshold setting. This project removed the data points in (1), and checked the distance of other data to the hyperplane in all LSSVMs, and compared the ratio of the distance from the support vector to the hyperplane with the set threshold (sigma). When there were two or more ratios that were smaller than sigma, then it was determined that the classification result of the main classifier was unreliable. At this time, the system selected K-D Tree for retest. Otherwise, it was determined that the main classifier could be trusted.

2.4 Algorithm model display and preprocessing instructions

In this paper, the above two designed and improved algorithms were combined to form a fault diagnosis algorithm design based on CS-optimized LSSVM and kNN algorithm implemented by Kd-tree, as shown in Fig. 1. After obtaining the training data, the project first separated the various types of data and preprocess the data. The data was divided in the "one-to-many" manner and undersampled to cooperate with the CS algorithm to train multiple LSSVMs and the building of Kd-tree classifier. The final output was judged by the primary and secondary classifiers. So far, the fault

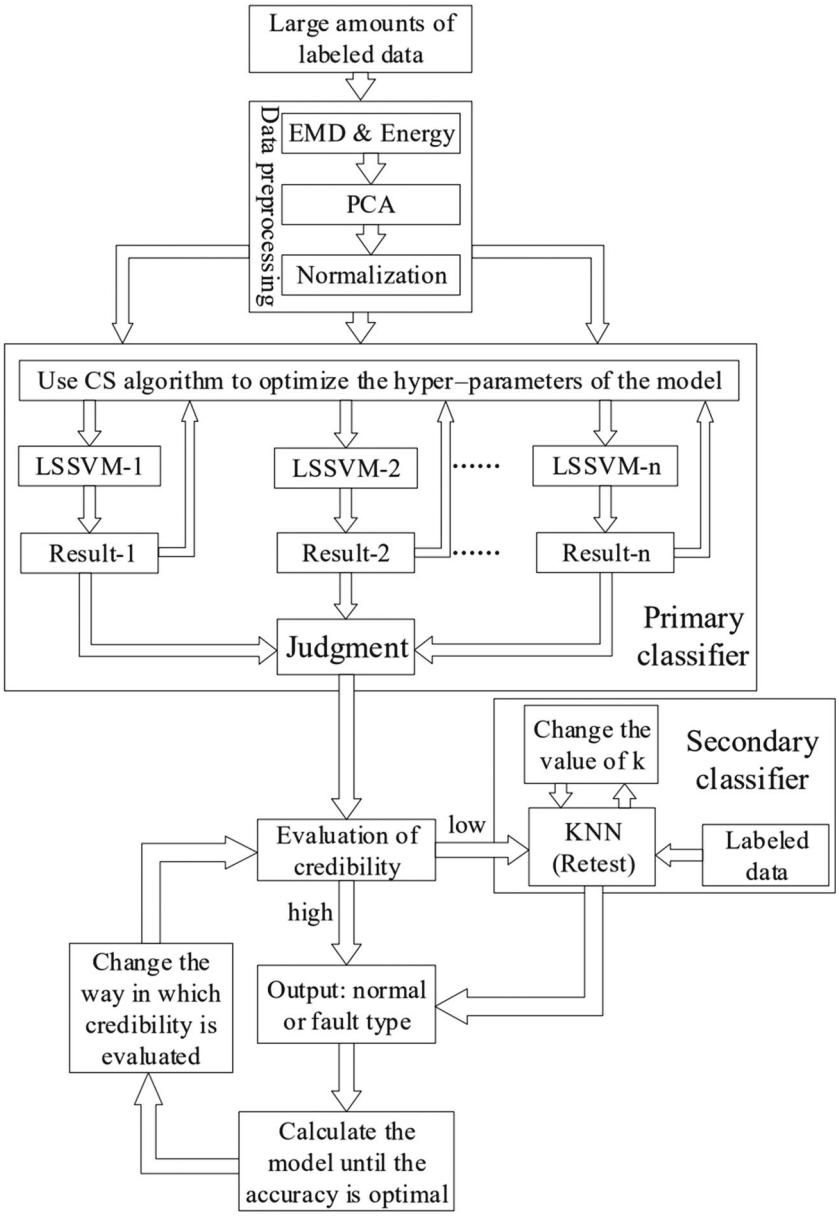


Fig. 1 Design of fault diagnosis algorithm based on LSSVM and Kd-Tree

diagnosis algorithm based on LSSVM and K-D tree was constructed.

In the data preprocessing, three processes including Empirical Mode Decomposition (EMD) & energy ratio calculation, Principal Component Analysis (PCA) and data normalization were adopted. Data normalization was performed using extreme value normalization.

Analysis of data over a period of time can more effectively extract signal features. Empirical Mode Decomposition (EMD) is widely used for bearing fault detection and has been proven to have good detection capabilities [33, 34]. EMD decomposes signals according to the time scale characteristics of the data itself, and decomposes a complex multi-component modulated signal into a set of single-component modulated signals in order from high to low frequency, that is, a set of intrinsic mode functions (imfs) [35–37]. The relationship before and after the decomposition is as in formula (4).

$$\xi(t) = \sum_{i=1}^{K} imf_i(t) + r_K(t) \tag{4}$$

Where, $\xi(t)$ is the original signal, $imf_i(t)$ is the decomposition result of the i-th layer. $r_K(t)$ is the residual after decomposition, which represents part of the amplitude information of the original signal.

Multi-layer data obtained by decomposing the original data by EMD is used to calculate energy according to formula (5).

$$E_{ck} = \left| \int C_k(t) \right|^2 dt = \sum_{j=1}^{n} |C_{kj}|^2 \tag{5}$$

Where $C_{kj}(j = 1, 2, \cdots)$ represents the signal amplitude of the j-th discrete point decomposed by the k-layer. Calculate the quotient with and the total energy to calculate the energy proportion of each layer [38, 39]. Different fault features have different energy distributions in each layer, so a certain number of layers are used to calculate the energy proportion of each layer, and the results can be used for fault diagnosis.

After the energy ratio results of multiple sensors were combined, the data dimension was still very large. In order to reduce the classification pressure, it was necessary to reduce the dimension to facilitate fault diagnosis. PCA uses linear transformation to transform the original data into variables that are linearly independent in each dimension, in order to extract the main feature components of the set of data, while effectively removing noise and redundancy [40].

Extreme value normalization is done as follows:

$$x_{ik} = \frac{x''_{ik} - x''_{ik\,\min}}{x''_{ik\,\max} - x''_{ik\,\min}} \tag{6}$$

Where, $x''_{ik\,\max}$ and $x''_{ik\,\min}$ are the maximum and minimum values in the k-th column parameters, and x''_{ik} is the parameter for the i-th and k-th columns of the faulty data set.

2.5 Prior clustering method designed for fault diagnosis

When the fault data was not clearly classified, in order to solve the impact of noise and outliers on the classification results and avoid the lack of prior knowledge caused by unknown fault categories, this project used FCM clustering to provide fault diagnosis algorithms with prior knowledge of fault data. Firstly, all fault data were aggregated into C categories, and then the C categories of fault data and non-fault data were used for training to reduce unknowns. These works reduced the error and impact of unknown priors on classification, transformed the two-classification task into a multi-classification task, and enhanced the practicality of the fault diagnosis algorithm.

In Fuzzy C-Means clustering algorithm (FCM), the so-called fuzzy is the process of using uncertainty instead of determination to reduce the error in a multi-probability form [41]. And use the clustering validity index λ_{MPC} to check the number of clusters C:

$$\lambda_{MPC} = 1 - \frac{C}{C-1}\left(1 - \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N} u_{ic}^2\right) \tag{7}$$

Where, u_{ij} is the degree of membership of sample j belonging to category I, C is the number of center points; N is the total number of samples.

The FCM algorithm has the disadvantage of slow clustering in principle. In the process of optimizing FCM in this paper, the particle swarm algorithm was used to assist FCM to find the number of clusters C faster.

Particle Swarm Optimization (PSO) is the research result of bird foraging behavior. Suppose a flock of birds is searching for food in a large area. When they do not know the location of the food, but know how far they are from the location, they can switch the search mode and perform the search by finding the closest individual to search for the food [42]. The PSO algorithm is inspired by such population behavior.

The implementation of particle swarm algorithm is simple, which can easily improve the accuracy of clustering and speed up the training speed of clustering. In this paper, PSO and FCM were combined, and then the FCM that incorporates the PSO algorithm (P-FCM algorithm) was obtained.

3 Execution steps

To sum up, in this paper the above two algorithms of LSSVM and K-D tree were merged to obtain the fault diagnosis algorithm design based on LSSVM and K-D tree. The implementation steps of the algorithm were given below.

3.1 Data preprocessing

(1) First, perform data completion processing on the missing samples.
(2) Format all the original data.
(3) Perform EMD decomposition and energy ratio calculation on the entire data set, take the first some layers with a cumulative energy ratio greater than 99% to calculate the energy ratios again, and combine the data calculation results of multiple sensors at the same time. Then apply the PCA data dimensionality reduction and data normalization to the above results for preprocessing.
(4) The binary classification problem uses P-FCM to provide prior knowledge of fault data.

3.2 Algorithm model display and preprocessing instructions

(1) The preprocessed data is extracted. The data is proposed to divide the data in a "one-to-many" manner and divide the training set / test set according to a

certain ratio, and prepare to train multiple LSSVM classifiers.

(2) Set CS algorithm parameters and stopping conditions. According to the requirements, it can be known that the CS algorithm's step size control ∂, and Levi index β are set to constants, $\partial = 0.1$, $\beta = 1.5$, the parameter of the probability of parasitic discovery P_a is initialized to 0.25; the initial number of bird nests n = 15. Set the maximum update generation number to 200 generations (number of iterations). Set the stop condition to the best fitness for 20 consecutive generations without change.

(3) Set the LSSVM hyper-parameter value range setting. According to design experience, $\gamma = C^{-2}$, $D^2 = 1/2\sigma^2$, thus $C \in [0.01, 10]$, $D \in [0.01, 50]$, in this design, the optimization of the regularization parameters of LSSVM and the width parameters of the kernel function is converted into finding and setting the parameter C and parameter D.

(4) CS contact LSSVM model. Use CS algorithm to select the transformed parameters and parameters of LSSVM, and then calculate the fitness (accuracy) on behalf of the vector machine. Save the current best fitness and the corresponding best bird nest, and calculate the average fitness and fitness after each update of the bird nest position.

(5) Determine whether the stop condition is met. If it does not arrive, continue to step (4). If it arrives, output the result parameter result. The LSSVMs independent algorithm models could be constructed by using the parameter optimization results and training set information.

3.3 Training of K-D tree based sub-classifier

When constructing the kNN model, Euclidean distance was selected as the distance calculation method, and weighted voting was used as the voting method. To prove the superiority of Kd-Tree performance, the data set is used to compare the performance of K-D tree with linear scanning.

When constructing the secondary classifier, the key is the selection of the parameter k. The optimal value of k is traversed in the range of [1,100].Take the pre-processed original fault data and non-fault data as input, the following steps are performed to construct a Kd-Tree:

(1) Construct the root node. Select $x^{(1)}$ as the coordinate axis, sort all training set samples on this axis, and then find the median node to divide all sorted training sets into two segments. The left and right child nodes are divided according to the relationship between the value of samples and the value of the median node.

(2) Iterative segmentation. The nodes $x^{(l)}$ are continuously used to segment the data of the j-th layer, and the segmentation rules satisfy the formula $l = j \,(\mathrm{mod}\, k) + 1$. In this way, a complete binary tree can be obtained.

(3) Termination. Repeat step (2) continuously until the sample points in the data set do not exist in the scope of the parent node's subordinate nodes, thereby obtaining a K-D tree.

3.4 Algorithm fusion of primary and secondary classifiers

After the above steps, multiple LSSVM classifiers and K-D tree classifiers could be obtained. Integrate the classification algorithm according to the following steps:

Main classifier fusion. The 4 trained LSSVMs are fused in a "one-to-many" manner.

Fusion of primary and secondary classifiers. (1) KNN retest. The primary and secondary classifiers are fused in the first way above to implement retesting. (2) Sigma threshold setting. On the existing data set, the threshold sigma is traversed between [0,1] in steps of 0.01 to test the fusion effect.

When the system collected new data, it preprocesses the newly input data, and uses the classification model and outputs the results.

3.5 P-FCM implementation

When fault data is not clearly classified, P-FCM is introduced to provide prior knowledge of the data. Specific steps are as follows:

(1) Import the data set to be processed.

(2) Parameter setting. Such as FCM hyperparameter C value range [2,100]. PSO acceleration parameters $C_1 = C_2 = 2$, and set the number of iterations to 100, and population number is set to 20. PSO stop conditions are not optimized for 10 generations.

(3) Initialization. The particle swarm is used to initialize the number of clusters in the FCM algorithm, and the fitness of each particle and the global optimal fitness are obtained according to the clustering validity index. The current optimal position of the particles can be compared to obtain the global optimal position.

(4) Use the optimization idea of particle swarm optimization to optimize the number of clusters of FCM algorithm. The particle fitness is obtained by calculating the clustering validity index, and the local and global optimal fitness update parameters are compared.

(5) Check the PSO stop conditions. If the PSO termination condition is not reached, the loop is executed (4).

When the PSO termination condition is reached, the current global optimal solution is retained.

4 Experiments

In order to verify the reliability of the overall design and the effectiveness of the algorithm, in this paper, multiple sensors including vibration, temperature, and pressure were installed on the electric fan of a power plant, and the data was collected at the same time and converted to PC monitoring software. Some data was extracted for algorithm model construction and testing, including 4 types (1 type of normal and 3 types of failures).

4.1 Data preprocessing

At the adoption frequency of 12k, 2048 time series data of a single sensor were taken as a group for EMD decomposition, and the decomposition results of one sensor data of the normal category are shown in Fig. 2.

The energy ratios of the EMD decomposition results were calculated. The first few layers of imfs were selected to recalculate the energy ratios according to the standard that the sum of energy ratios exceeded 99% and the selected imfs were easy to handle. Finally, the number of imfs layers of sensor 1 under four types was 6, the number of imfs layers of sensor 2 is 6, and the number of imfs layers of sensor 3 is 7. In this case, the energy ratios is shown in Table 1.

The above-mentioned preprocessing results of three processor data of the same category and the same time were spliced to obtain 19-dimensional data, which was subjected to PCA data dimensionality reduction processing, and the result is shown in Fig. 3. As shown in the figure, as the dimensions of the data samples retained by PCA rise, the retention ratio of the original data's feature information also gradually increases. When the number of data dimensions after the dimension reduction is 5, the retained information ratio reaches 99.62%; it is 6 at that time, the retention ratio reaches 99.88%; when it is 7, the retention information ratio reaches 99.96%.

Table 1 The proportion of the sum of the extracted imfs' energy

Category	Normal−0	Fault−1	Fault−2	Fault−3
sensor 1	99.09%	99.12%	99.03%	99.25%
sensor 2	99.66%	99.17%	99.26%	99.33%
sensor 3	99.54%	99.43%	99.52%	99.10%

In order to reduce the information loss to less than 0.1%, the number of dimensions of the data after dimension reduction is selected to 7, which reduced the interference of some unrelated features on the training process and speeded up the training speed.

The data after dimensionality reduction was normalized according to the extreme value normalization method mentioned above. Extreme values of each dimension are shown in Table 2. At this point, the preprocessing process was completed. The data was divided into training set and test set, and the distribution of data volume was shown in Table 3.

4.2 Construction of LSSVM models optimized with CS

The preprocessed 4 types of data were divided in a "one-to-many" manner and appropriately under-sampled. The training results of the four CS-optimized LSSVMs trained on this data set are shown in Figs. 4, 5, 6, and 7.

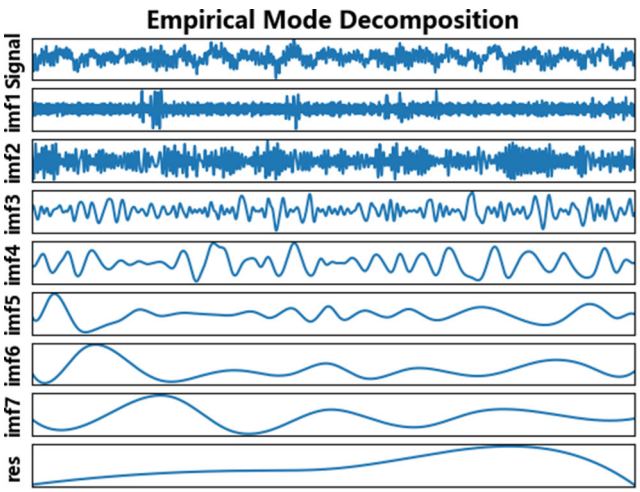


Fig. 2 The decomposition results of a sensor in the normal category

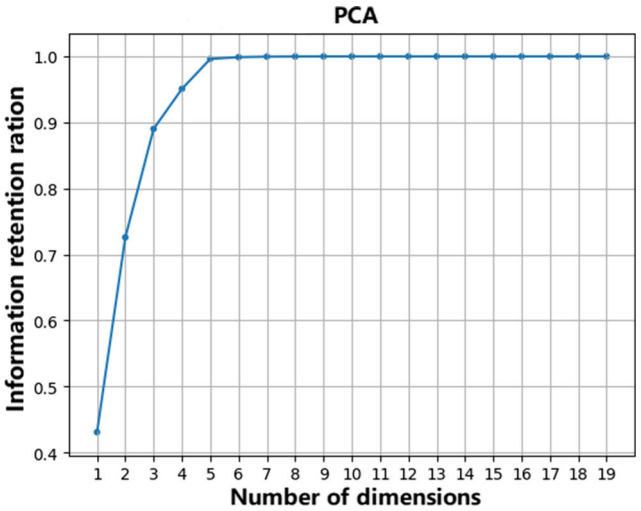


Fig. 3 Relationship between PCA algorithm dimension and information retention

Table 2 Extreme value table for each dimension of the sample data

Dimension	0th	1st	2nd	3rd
Max	1459.3385365408	147.2652785376	154.1940114499	558.0688314725
Min	−24.4727774317	−133.7409962635	−135.5679861337	−339.0044860397
Dimension	4th	5th	6th	——
Max	122.1258465574	40.2882781031	31.5363908832	——
Min	−101.1718470581	−16.2546276676	−28.9255085563	——

The optimal parameters of the 4 LSSVMs are shown in Table 4.

4.3 Optimization and implementation of kNN classifier

The performance of Kd-Tree and linear scan were compared as described in Section 3.3. A data set with 7,400 20-dimensional data samples was used to compare the two algorithms (the data set had two types, 3664 for class 0 data and 3736 for class 1 data), and preprocessing was required before the data was used. The following three aspects were compared respectively:

(1) Randomly extracted 4000 data (3000 training data, 1000 test data, and reduced the data dimension to 8 dimensions), and the value of k was taken as [1,100]. The changes of the accuracy of the two are shown in Figs. 8 and 9 below.
(2) Under condition (1), the changes of the time consumption of them with the change of k are shown in Figs. 10 and 11.
(3) 1000 test data were taken; k = 5; the data dimension was 8 dimensions; the size of training data was traversed in steps of 60 within the range of [60,6000]. The changes of the time consumption of them with the change of the data size are shown in Figs. 12 and 13.

It can be seen from the comparison in (1) that the Kd-Tree and the linear scanning method are the same in finding the nearest neighbors when implementing the kNN algorithm, which proves the accuracy and effectiveness of Kd-Tree.

It is easy to know from (2) that when the amount of training data is much larger than the spatial dimension, the time consumption of the Kd-Tree increases with the increase of the value of k, and the change in the time consumption of linear scan shows an unstable phenomenon. At the same time, it can be known that under the same conditions, the kd tree takes much less time than the linear scan, which proves that the kd tree is efficient in implementing the kNN algorithm while ensuring a certain amount of training data.

It is also easy to know from (3) that when the data size is much larger than the dimension, when the size of data increases linearly, the time consumption of the Kd-Tree increases slowly, and the time consumption of the linear scan shows a linear increase, and the time consumption of the Kd-Tree is much smaller than the linear scan . Therefore, the Kd-Tree is more efficient and real-time than linear scan.

In this paper, when constructing the secondary classifier, k was traversed from 1 to 100, and the variation of accuracy and time was shown in Figs. 14 and 15. As can be seen from the figure, when the value of k is set to be 6, the highest accuracy of the kNN algorithm model, 0.921687 (92.17%) can be reached and the search takes less time with the given data set of this paper. Embedded the value of k and the training data set into the kd tree and the construction of sub-classifier kd tree could be completed.

Table 3 Data volume distribution of training and test sets

Category	Normal−0	Fault−1	Fault−2	Fault−3
Training set	247	247	247	247
Test set	82	84	83	83

4.4 Algorithm fusion and testing of primary and secondary classifiers

Multiple classification algorithms were fused step by step. The sigma threshold in Section 3.4 above traverses Fig. 16.

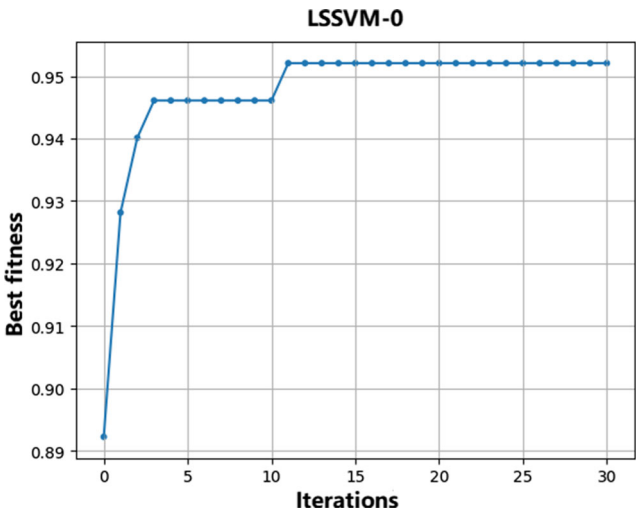


Fig. 4 Best fitness change of LSSVM-0

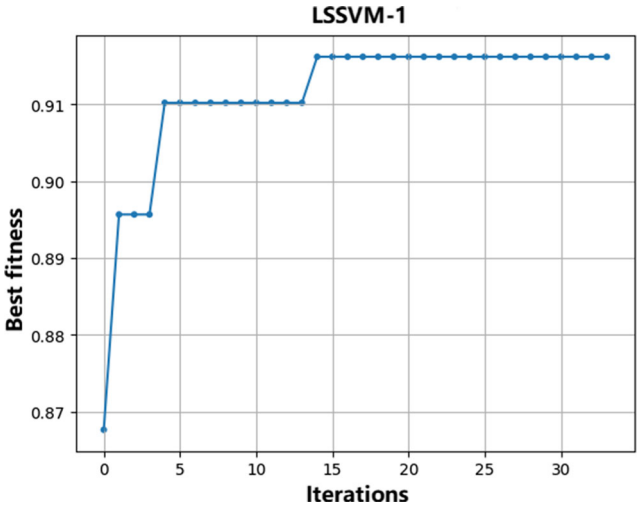


Fig. 5 Best fitness change of LSSVM-1

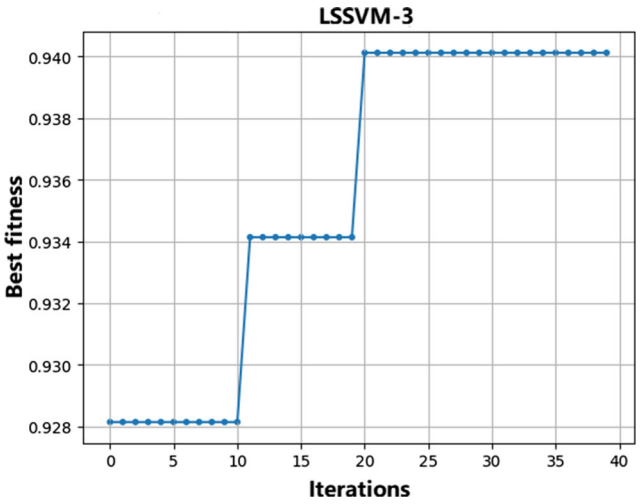


Fig. 7 Best fitness change of LSSVM-3

In order to improve the efficiency and accuracy of the overall algorithm, the sigma should be taken as small as possible when the overall accuracy is the highest. Therefore, the sigma was selected as 0.54.

After the algorithm model was constructed, the overall test was performed. The fusion of the steps and the overall effect are shown in Table 5.

In traditional multi-classification tasks, a single multi-classification algorithm or a two-classification algorithm is often used to complete the multi-classification task. As can be seen from Table 4 and Figs. 4 to 7, LSSVM performs very well when it completes the binary classification tasks. As can be seen from Table 3, under the conditions of this paper, when using LSSVM to complete the multi-classification task, the accuracy rate is only 87.65%. Therefore, the accuracy of LSSVM decreases when multi-classification tasks are completed, which proves the necessity of introducing a secondary classifier.

In this paper, the primary and secondary classifiers were fused, which greatly improved the overall accuracy of the algorithm model. Finally, the accuracy of the overall model reached 95.18%, thereby proving the rationality and effectiveness of the algorithm model.

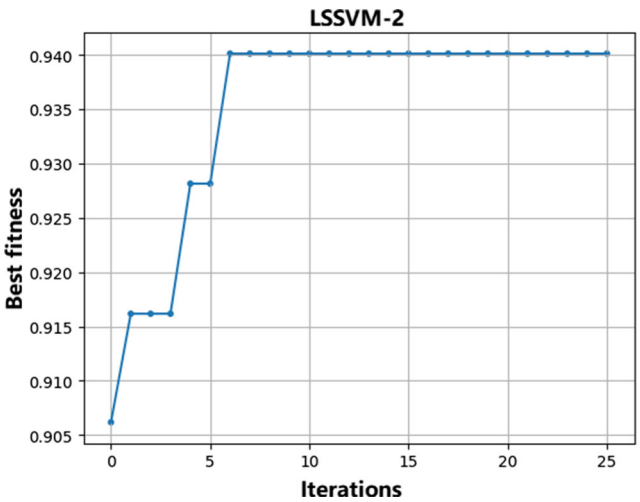


Fig. 6 Best fitness change of LSSVM-2

4.5 P-FCM algorithm experiment

The fault data of the data used in this paper was mixed, and the prior knowledge was provided by P-FCM. The result is shown in Fig. 17 (the number in the figure represents the corresponding parameter C under the best clustering). Using ordinary FCM, that is, the method of traversing the parameter C, the result is shown in Fig. 18. The optimal number of C is 3. PSO updated 2 generations to find the optimal value, which is consistent with Fig. 16 and the clustering accuracy is 76.67%. Because there were not many data categories in this paper, PSO didn't show great search advantages. In large samples, the rapid characteristics of PSO could be reflected.

The mixed fault data and normal data were classified by LSSVM, and the accuracy rate was 90.96%. The accuracy rate based on the P-FCM prior was 92.77%. The effectiveness of introducing the P-FCM algorithm to the

Table 4 Model parameter selection and accuracy of LSSVMs

Model	Parameter C	Parameter D	Iterations	Accuracy
LSSVM−0	0.0327284228	3.1629143553	11	95.21%
LSSVM−1	0.3411471486	5.5421693265	14	91.62%
LSSVM−2	1.1257903231	0.0486292310	6	94.01%
LSSVM−3	0.0388663249	0.3081453070	20	94.01%

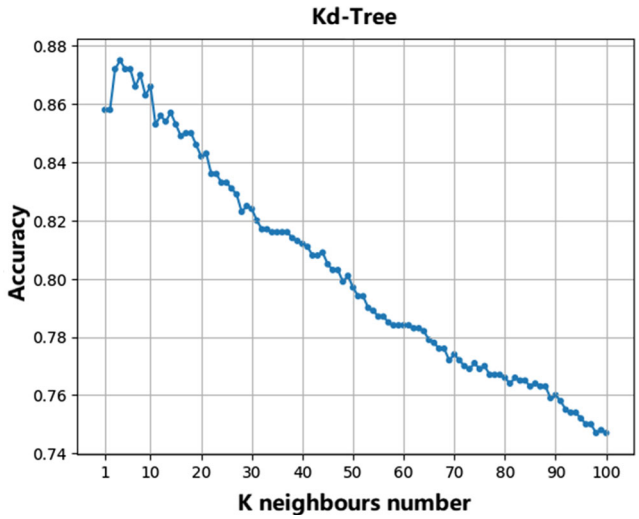


Fig. 8 Kd-tree accuracy curve as parameter k changes

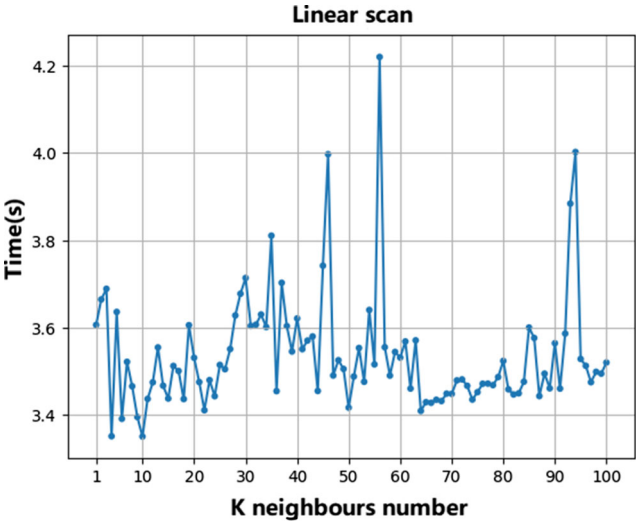


Fig. 11 Time-consuming change of linear scan

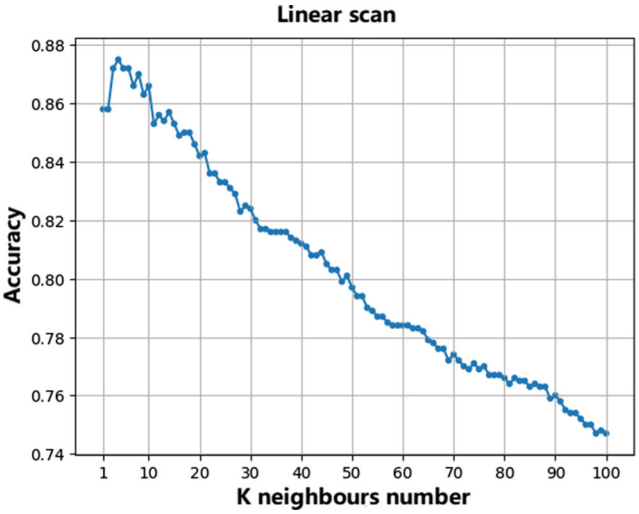


Fig. 9 Linear scan accuracy curve as parameter k changes

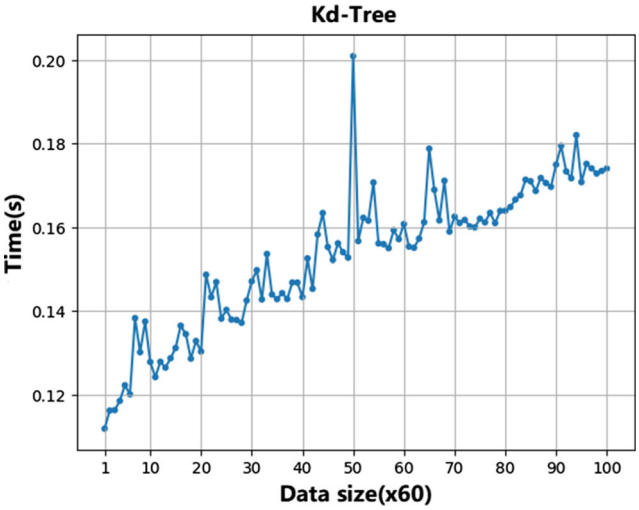


Fig. 12 Time-consuming change of Kd-Tree

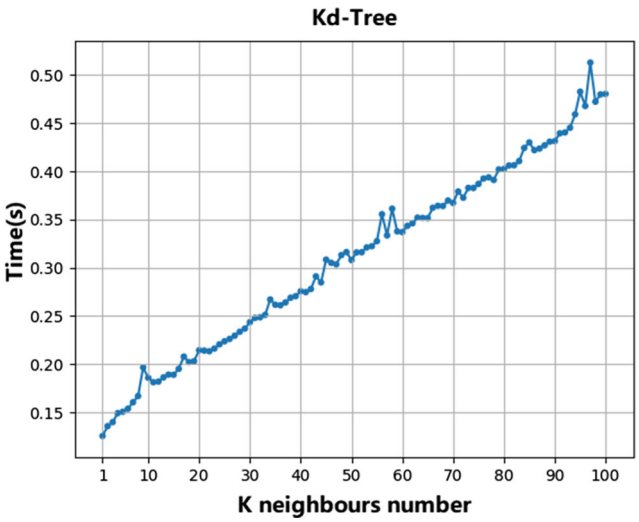


Fig. 10 Time-consuming change of Kd-Tree

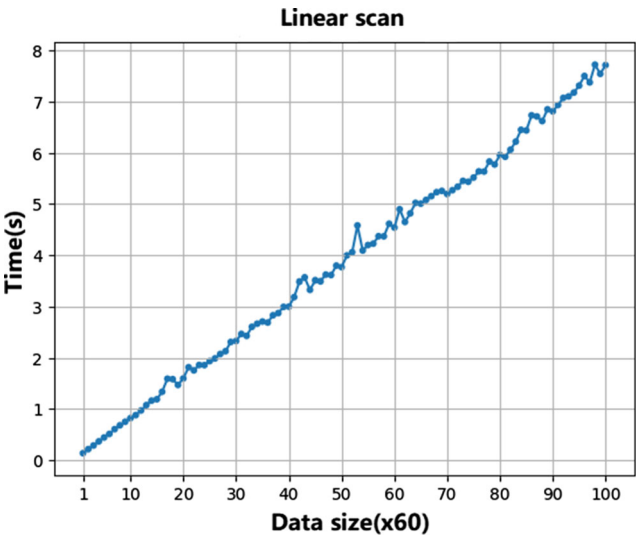


Fig. 13 Time-consuming change of linear scan

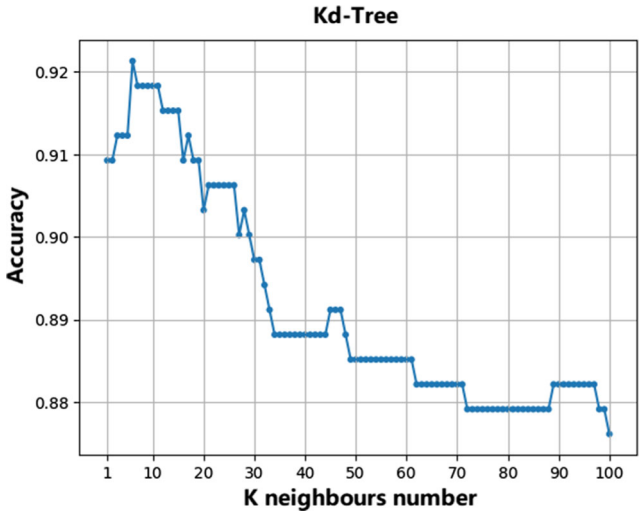


Fig. 14 Accuracy curve of the secondary classifier

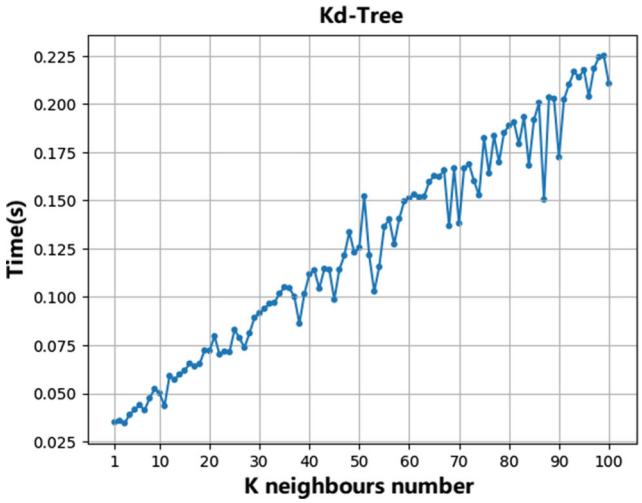


Fig. 15 Time consumption curve secondary classifier

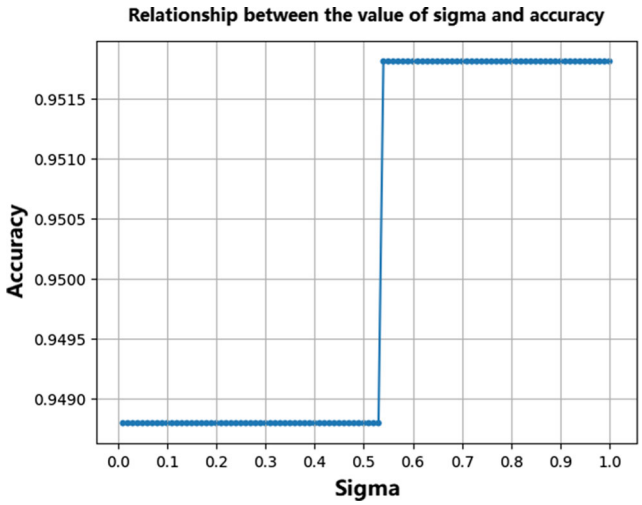


Fig. 16 Relationship between the value of sigma and accuracy

Table 5 Primary and secondary classifiers fusion effect display

Primary classifiers	KNN retest	Sigma threshold setting	The entire model
Accuracy: 87.65%	Accuracy increased by 7.23%	Accuracy increased by 0.30%	Accuracy: 95.18%

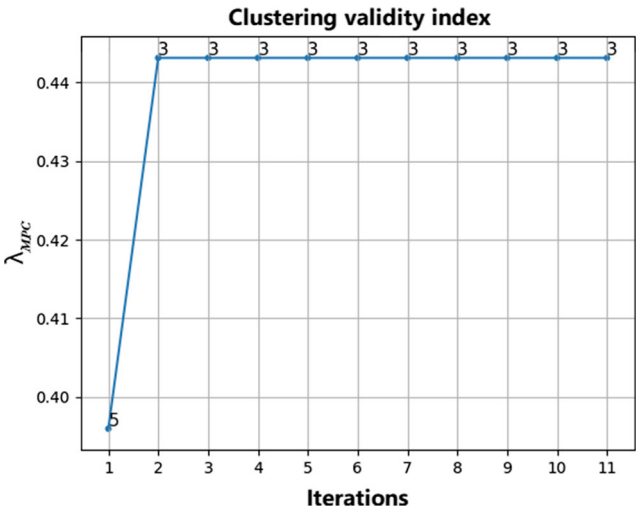


Fig. 17 P-FCM calculation results

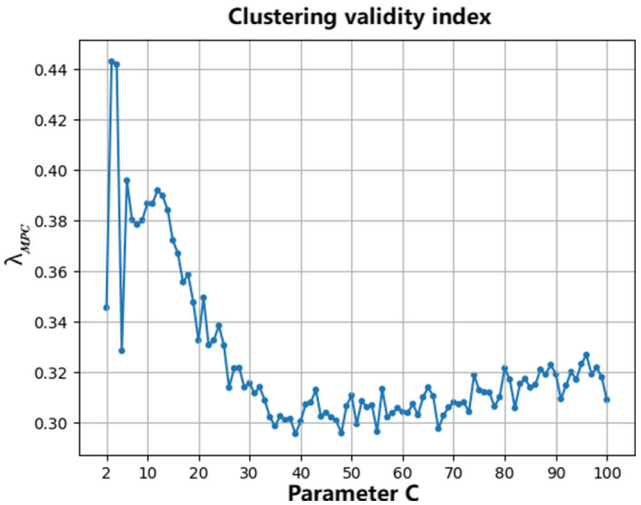


Fig. 18 Ordinary FCM operation results

Table 6 Classifier performance comparisons table

Model	Naive Bayes classifiers	Logistic Regression	Decision Tree	BP Neuron Network	LSSVM	Kd−Tree	My model
Time(s)	$4.71e-5$	$1.68e-4$	$1.38e-4$	$2.73e-4$	$1.79e-4$	$6.02e-5$	$1.91e-4$
Accuracy	78.01%	89.76%	90.96%	94.28%	87.65%	92.17%	95.18%

binary classification problem was proved. Therefore, the introduction of P-FCM algorithm could provide a priori knowledge for the model, which expanded the practicality of the fault diagnosis algorithm model in this paper, and proved the validity and rationality of the model from the side.

4.6 P-FCM algorithm experiment

In this paper, the primary and secondary classifiers were combined in the above manners, and the performance comparisons of this model and other commonly used algorithms were performed on the pre-processed data set shown in Table 4. The algorithms involved in the comparisons included Naive Bayes classifiers, Logistic Regression, Decision Tree, 3-layer BackPropagation Neuron Network (BP Neuron Network) (hidden neurons are 100, 50, 20) etc. The results are shown in Table 6. The time in the table is the average detection time of the algorithm for a single sample, in seconds (s).

From the comparison results in Table 6, it can be seen that the algorithm model formed by the fusion of the primary and secondary classifiers proposed in this paper performs better than other commonly used single models in accuracy. Although the model used two classifiers, the Kd-Tree as a sub-classifier was only used for partial data re-examination. And the algorithm efficiency of the Kd-Tree is relatively high. Therefore, the average detection time of this model is not significantly worse than other algorithm models, and it is faster than BP neural network according to Table 6. It could be seen that the model showed better performance after a compromise between algorithm speed and accuracy.

4.7 Other relevant data processing experiments

When the original experimental data of the project was screened, special attention has been paid to the problem of data balance, as shown in Table 3. But the problem of unbalanced data was still encountered when constructing the main classifier. In order to reduce its impact, this project used the processing method of undersampling. Under this condition, the performance of the algorithm refers to (3) in Table 7.

However, undersampling inevitably loses part of the data information. So the GAN is used for additional experiments. Because the problem of data imbalance in constructing the main classifier was relatively light, the network structure of GAN used was relatively simple. And its generator and discriminator were all composed of three hidden neurons. The hidden neurons of the generator were 128, 256, and 512, and the hidden neurons of the discriminator were 100, 50, and 20, respectively. The data after processing was substituted into the model of this paper for testing. And the results are shown in (1) of Table 7.

The main classification of this subject was composed of 4 LSSVMs, and the efficiency inevitably decreased. In this paper, in view of this problem, considering that the accuracy performance of LSSVM is related to support vectors, half of the data closer to the support vector of LSSVM in the data was selected on the basis of undersampling, and the balance of data was paid attention to in this process. Put the screened data into the model of this paper for testing, the results are shown in (2) of Table 7. The time in Table 7 is the average detection time of the algorithm for a single sample, in seconds (s).

It can be seen from (1) and (3) of Table 7 that the GAN's processing method of data imbalance has an improvement effect on the model accuracy of this subject, but the ensuing increase in model calculation time. According to (2) and (3) in Table 7, by selecting the data closer to the support vector, the model running time can be reduced, but the accuracy of the model is also affected. Therefore, it can be seen that the algorithm combined with undersampling proposed in this paper is a compromise between the accuracy and efficiency of the model, and has excellent performance compatible with accuracy and efficiency. The other two data processing

Table 7 Comparison of classifier performance under different data processing methods

Number	Processing method	Time(s)	Accuracy
(1)	GAN	$2.82e-4$	95.78%
(2)	Undersampling +Selection	$1.79e-4$	93.37%
(3)	Undersampling	$1.91e-4$	95.18%

methods introduced in this section are complementary to the application scenarios of the overall algorithm model in this paper, which expands the universality of the model in this paper.

5 Summary and work prospects

With the widespread use of machinery and equipment in the industry, machinery equipment fault diagnosis has attracted wide attention. Based on the existing monitoring and diagnosis system of electric fans, this paper designed an algorithm model based on LSSVM and K-D tree algorithm.

Based on data preprocessing, the model built a main classifier based on CS-optimized LSSVM, and introduced a KNN algorithm based on Kd-Tree to improve performance. The primary was fused based on the "one-to-many" method of the data. In addition to the kNN retest proposed on this basis, the primary and secondary classifiers also introduced a sigma threshold judgment fusion method to improve the overall algorithm model performance. After the algorithm test, the accuracy of the fault diagnosis algorithm designed in this paper reached 95.18%, which was 7.53% higher than that of the LSSVM primary classifier. The kNN retest contributed 7.23% to the accuracy improvement, while the contribution of sigma threshold was 0.3%. Besides, in view of the lack of prior knowledge, a FCM optimized by the PSO algorithm, namely P-FCM, was proposed to provide prior knowledge; transformed the binary classification problem into a multi-classification one; reduced the errors caused by the diversified distribution of fault data; and made the classifier fault diagnosis algorithms become more widely used. Compared with a single LSSVM, the accuracy of this algorithm was improved by 1.81%. Comparing the performance of the algorithm model based on LSSVM and K-D tree with other commonly used models in this paper, we could see that the model had good algorithm performance. The introduction of the other two data processing methods and the relevant experiments expanded the universality of the algorithm.

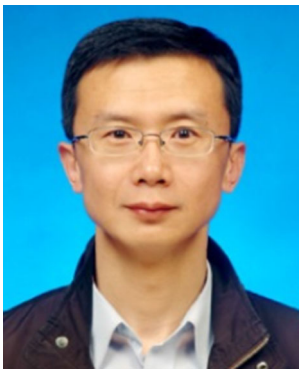
In the future, the algorithm proposed in this paper should be experimented with more different types of data sets and the algorithm structure should be modified. Apart from that, the P-FCM algorithm should be further optimized and implemented so that the proposed algorithm can be stronger and adaptable.

References

1. Qian W (2019) Key causes and diagnosis technology of automobile machinery faults [J]. Shandong Ind Technol 15:38
2. Song Y, Wu K, Chu N, et al. (2019) Research on common fault diagnosis methods of subway fans based on modulation intensity [J]. Fan Technol 1:77–81
3. Gao ZW, Cecati C, Ding SX (2015) A survey of fault diagnosis and Fault-Tolerant techniques-Part i: fault diagnosis with Model-Based and Signal-Based approaches [J]. IEEE Trans Ind Electron 62(6):3757–3767
4. Li YQ, Yang TS, Liu J, et al. (2016) A fault diagnosis method by multi sensor fusion for spacecraft control system sensors[C]. In: 2016 IEEE international conference on mechatronics and automation, pp 748–753
5. Cui L, Huang J, Zhang F (2017) Quantitative and localization diagnosis of a defective ball bearing based on vertical-horizontal synchronization signal analysis. IEEE Trans Ind Electron 64(11):8695–8705
6. Wang J, Du G, Zhu Z, et al. (2020) Fault diagnosis of rotating machines based on the EMD manifold[J]. Mech Syst Signal Process 135:1–21
7. Kinkead N (1985) Structural integrity monitoring: R.A. Collacott. Chapman and Hall, London. 488 pp, Â£49.50[J] 8(4), 1986
8. Lei Y, Zuo MJ (2009) Gear crack level identification based on weighted K nearest neighbor classification algorithm[J]. Mech Syst Signal Process 23(5):1535–1547
9. Ali MZ, Shabbir MNSK, Liang X, et al. (2019) Machine Learning-Based Fault Diagnosis for Single- and Multi-Faults in Induction Motors Using Measured Stator Currents and Vibration Signals[J]. IEEE Trans Ind Appl 55:2378–2391
10. Fu W, Tan J, Zhang X, et al. (2019) Blind parameter identification of MAR model and mutation hybrid GWO-SCA optimized SVM for fault diagnosis of rotating machinery. Complexity 2019:1–17
11. Shi Z, Ge C (2018) Turbine rotor fault diagnosis based on CS-BBO optimized SVM [J]. Vibr Test Diagn 38(03):619–626
12. Chen L, Zhang Z, Cao J, Wang X (2020) A novel method of combining nonlinear frequency spectrum and deep learning for complex system fault diagnosis[J]. Measurement 151:1–8
13. Yang B, Zio E, Liu R, et al. (2018) Artificial intelligence for fault diagnosis of rotating machinery: A review[J]. Mech Syst Signal Process 108:33–47
14. Huang Y, Huang Y, Huang R (2018) Gear fault diagnosis based on BP neural network[J]. IOP Conf Ser Mater Sci Eng 322(7):072043 (5pp). https://doi.org/10.1088/1757-899X/322/7/072043
15. Huo L, Zhang X, Li H (2018) Bearing fault diagnosis based on BP neural Network[J]. Iop Conf 208:012092
16. Guo L, Li N, Jia F, Lei Y, Lin J (2017) A recurrent neural network based health indicator for remaining useful life prediction of bearings. Neurocomputing 240:98–109
17. Cartella F, Lemeire J, Dimiccoli L, Sahli H (2015) Hidden Semi-Markov Models for Predictive Maintenance [J]. Mathematical Problems in Engineering 2015:1–23
18. Comon P, Jutten C (2010) Handbook of blind source separation independent component analysis and applications. Informatica 35:824
19. Jia F, Lei Y, Lin J, Zhou X, Lu N (2016) A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data[J]. Mech Syst Signal Process 72–73:303–315
20. Mirsamadi S, Hansen JohnHL (2019) Multi-domain adversarial training of neural network acoustic models for distant speech recognition[J]. Speech Comm 106
21. Gao W (2018) Research on power fault diagnosis method of pumping unit based on hybrid intelligent technology [J]. Computer and Digital Engineering 046(009):1905–1910
22. Luo R (2019) Analysis of condition monitoring and fault diagnosis methods for electrical equipment in power plants [J]. Shandong Ind Technol 14:159

23. Wu C, Zhu C, Sun J, et al. (2017) A synthesized diagnosis approach for lithium-ion battery in hybrid electric vehicle[J]. IEEE Trans Veh Technol 66(7):5595–5603
24. Zhao X, Ye Z (2007) Intrusion detection classification algorithm based on weighted multi-stochastic decision tree [J]. Computer Engineering and Applications 18:135–137
25. Pan C, Li W (2018) Application of prior knowledge in intelligent fault diagnosis of gears at different speeds [J]. Mechanical Transmission 42(11):152–158
26. Zhou F, Yang S, Fujita H, Chen D, Wen C (2020) Deep learning fault diagnosis method based on global optimization GAN for unbalanced data[J]. Knowledge-Based Systems 187:1–19
27. Duan X, Chen D, Fan X, et al. (2020) Research and implementation on power analysis attacks for unbalanced data
28. Dan Y, Zhao Y, Li X, et al. (2019) Generative adversarial networks (GAN) based efficient sampling of chemical space for inverse design of inorganic materials[J]
29. Bemani A, Xiong Q, Baghban A, Habibzadeh S, Mohammadi AH, Doranehgard MH (2019) Modeling of cetane number of biodiesel from fatty acid methyl ester (FAME) information using GA-, PSO-, and HGAPSO-LSSVM models[J]. Renewable Energy 150(1):924–934
30. Lan S, Liu S (2015) Review of cuckoo search algorithm research [J]. Comput Eng Des 4:1063–1067
31. Zhu D, Zhu H, Liu X, Li H, Wang F, Li H, Feng D (2020) CREDO: Efficient and Privacy-Preserving Multi-Level Medical Pre-Diagnosis Based on ML-KNN[J]. Information Sciences 514:244–262
32. Hu L, Nooshabadi S (2019) High-dimensional image descriptor matching using highly parallel KD-tree construction and approximate nearest neighbor search[J]. Journal of Parallel and Distributed Computing[J]. Journal of Parallel and Distributed Computing 132:127–140
33. Zhang CL, Yue X, Jiang YT, et al. (2010) A Hybrid Approach of ANN and HMM for Cutting Chatter Monitoring[J]. Advanced Materials Research 905:3225–3232
34. Xia J, Su T, Ma Z, Leng Y, Bai Y (2013) EMD-based rolling bearing fault feature extraction method [J]. Noise and Vibr Control 33(02):123–127
35. Mi S, Xu J, Ming W, Chen M, Chen L (2020) Online detection of chatter vibration of cylinder head based on EMD and SVM [J]. Tool Technol 54(02):74–77
36. Gao C, Wu T, Fu Z (2018) Fault feature extraction method of rolling bearing based on improved EMD [J]. Softw Guide 17(12):156–160
37. Jiang X, Cao J, Hu J, Xiong X, Liu J (2020) Pre-stack gather optimization technology based on an improved bidimensional empirical mode decomposition method[J]. Journal of Applied Geophysics 177:1–11
38. Shi Q, Xu X (2018) Using EMD energy ratio and GA-BP network to diagnose rotating machinery faults [J]. Noise Vibr Control 38(02):168–172
39. Gao J, Shang P (2019) Analysis of complex time series based on EMD energy entropy plane[J]. Nonlinear Dynamics 96:465–482
40. Kundu P, Darpe AK, Kulkarni MS (2019) Weibull accelerated failure time regression model for remaining useful life prediction of bearing working under multiple operating conditions[J]. Mech Syst Signal Process 134:1–19
41. Tang Y, Ren F, Pedrycz W (2020) Fuzzy C-Means clustering through SSIM and patch for image segmentation[J]. Applied Soft Computing 87:1–16
42. Zhang J, Tang L, Liao B, Zhu X, Wu F-X (2019) Finding Community Modules of Brain Networks Based on PSO with Uniform Design[J]. BioMed Research International 2019(3):1–14

Kongzhi Hu received the Bachelor of Engineering from Harbin Institute of Technology in 2019. He is currently studying for the Master of Electronic Science and Technology from the School of Astronautics of Harbin Institute of Technology. His main research and learning directions include machine learning and computer vision.



Ming Jiang received the Master degree in signal and information processing from the School of Astronautics of Harbin Institute of Technology. And he received Bachelor degree in computer and application from the School of Computing of Harbin Institute of Technology. Presently, he is working as a teacher in the School of Astronautics of Harbin Institute of Technology. His current research interests include wireless sensor networks, signal processing and machine learning.

Haifeng Zhang received Ph.D. degree in Microelectronics and Solid State Electronics from Harbin Institute of Technology. And he received the Master degree in signal and processing from the Air Force Engineering University. The current title of Dr. Zhang is associate professor, and he is currently a doctoral supervisor in the School of Astronautics of Harbin Institute of Technology. He has published more than 30 SCI papers, obtained 17 authorized Chinese national invention patents, and served as a reviewer for several international journals. Dr. Zhang is also an IEEE member and a senior member of the Chinese Institute of Electronics. His current research interests include wireless sensor networks, signal processing and MEMS technology, etc.

Sheng Cao received the Master of Electronic Science and Technology from Harbin Institute of Technology in 2019. He is currently engaged in computer vision related work in a research institute. His main research areas include computer vision and data mining.

Ziyi Guo will receive the Bachelor of Engineering in Electronic Science and Technology from Harbin Institute of Technology in 2021. She is currently researching as an undergraduate intern at DASLab in Harvard University. Her research interest includes data processing and machine learning.