ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20230505000608
Original scientific paper

Health Prognosis for Equipment Based on ACO-K-Means and MCS-SVM under Small Sample Noise Unbalanced Data

Qinming LIU, Fengze YUN, Ming DONG, Darko DJORIC, Nikola ZIVLAK*

Abstract: For the problem of manufacturing system residual life prognosis with insufficient small sample data and unbalanced distribution, this paper proposes a model for equipment health status analysis and life prognosis based on improved ant colony optimization K-Means (ACO-K-Means) and multi-classification Self-Adding SVM (MCS-SVM). First, based on the fuzzy data set, the data is classified for the first time according to the traditional SVM, and the initial classification results are obtained. Second, the improved K-Means algorithm based on the ant colony algorithm is used to cluster the data set after the initial classification, to obtain more health status labels in different states.The noise scale coefficient is established, and the data set distribution is optimized by introducing the unbalanced scale standard and the adaptive addition rule, to enrich the sample capacity of the scarce label under the influence of noise. On this basis, the SVMset is introduced according to the number of clusters to achieve multi-classification of the data set. Finally, by using the state data of the hydraulic pump of Caterpillar, the simulation results show that the two improved algorithms can accurately analyze the health state and lifetime prognosis of equipment under small noise samples and unbalanced data.

Keywords: health prognosis, machine learning algorithm Noise data, state recognition, unbalanced data

1 INTRODUCTION

With the continuous development and progress of modern technology, the accuracy of equipment life assessment in industrial production is getting higher and higher. Timely and effective assessment of equipment health status and prediction of equipment remaining life level are directly related to the work efficiency of enterprises. Once people's assessment of the equipment deviates, it is likely to cause huge economic losses and casualties [1]. Thus, the effective evaluation and health management of machinery and equipment has become the focus of more and more scholars.

Compared with the traditional equipment life prediction scheme, the equipment health assessment based on automatic machine learning algorithm has gradually become the mainstream of modern enterprise research. The machine learning algorithm for equipment data with existing status tags is called supervised learning [2]. The machine learning algorithm for unfamiliar data without status tags is called unsupervised learning [3]. For supervised learning, a series of scholars have proposed various improvement schemes based on traditional machine learning. Among them, Garrido et al. [4] proposed an improved SVM based on PSO. By optimizing the dynamic weight, SVM can process nonlinear data, and then use the efficient optimization method of PSO to determine parameters to maximize the classification effect of the improved SVM. Guha et al. [5] proposed an optimization algorithm based on a variety of Bayesian combined neural networks, and improved the prediction accuracy of the algorithm through the evaluation of parameters.

In the field of unsupervised learning, Huang et al. [6] established an optimization algorithm for feature knowledge transfer and achieved good results. Huang et al. [7] established a machine learning model based on independent forests to label data sets and achieved good results. Kwon et al. [8] proposed an unsupervised fault diagnosis method that can realize fault detection and location based on the normal sample training classification model. McLeay et al. [9] adopted the unsupervised learning method for equipment fault detection and proved the effectiveness.

For unbalanced data, scholars have also carried out some research in this field. Among them, Li et al. [10] proposed using the global information addition method to add a few valuable sample points and achieved good results. Liu et al. [11] combine under sampling and oversampling when facing unbalanced data, and use voting random forest to enhance the classification effect. Liu et al. [12] proposed a rolling bearing life stage recognition method based on multi-classifier integration and weighted balanced distribution adaptation. Liu et al. [13] solved the problem that the single body could not be tested due to unbalanced data in the field of box fault diagnosis and life prediction. Lv et al. [14] adopted the SMOTE algorithm to process the unbalanced data and successfully predicted the health of the optical cable. Li et al. [15] proposed a unified framework and model for fusion prediction to generate de-noising self-coder and deep coral network.

Under the noise data, Peng et al. [16] established an improved combination model based on support vector machine for the noise data to predict the residual life of bearings. Wu et al. [17] established a support vector machine prediction model based on noise reduction data. Yan et al. [18] proposed a life prediction method of hydraulic cylinder based on deep learning. The noise data is reconstructed using the DAE algorithm, avoiding the problems that may be caused by the noise data, and the life of hydraulic cylinder is successfully predicted. Bhouri et al. [19] and Maddu et al. [20] proposed a processing method for this kind of data, but it did not take into account the problems of sample strangeness and imbalance.

Scholars extensively research supervised and unsupervised learning for optimizing noise and unbalanced data. However, in practical industries, equipment life data obtained from testing often suffers from incompleteness, such as insufficient samples, ambiguous states, and uneven distribution. Existing studies focus on limited aspects of machine learning algorithm optimization, particularly in equipment life prediction. Few studies address challenges related to imbalanced equipment health data samples, small sample sizes, and fuzzy labels. To tackle these issues,

this paper investigates incomplete data in industrial production and proposes a model for small sample imbalance, fuzzy data, and noise. The simulation stage involves calculating and analyzing sample health status and predicting equipment residual life based on root mean value evaluation of equipment vibration. The aim is to simulate actual industrial data characteristics, enhance K-Means algorithm efficiency using ACO, introduce new rules and noise ratio rules for building an improved SVM set for classification, and ultimately achieve equipment status recognition and health prognosis.

2 HEALTH PROGNOSIS MODEL BASED ON IMPROVED ACO-K-MEANS AND MCS-SVM

2.1 Poblem Description

In the context of intelligent industrial production systems, equipment performance has improved, reducing failure probability. However, interconnections among components can cause system breakdowns if a single component fails. Efficient fault detection and prediction in low sample quantity and imbalanced samples are crucial in fault diagnosis and life prediction research. To tackle these issues, this paper uses various algorithms and models, enhancing classification performance of the samples. Fig. 1 illustrates the technical road map.

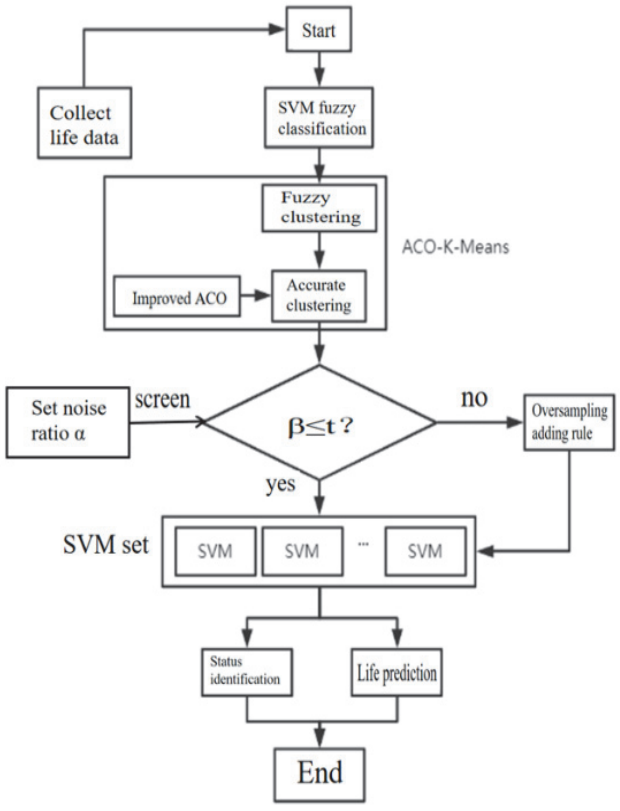

Figure 2 Technology roadmap based on ACO-K-Means and MCS-SVM algorithm

2.2 ACO-K-Means

In this paper, the core idea of improving K-Means is to limit the search range of the traditional algorithm using ACO, reducing complexity and enhancing clustering effectiveness. The improved algorithm employs fuzzy clustering, where the ant colony algorithm's search scope is determined based on the results of the initial fuzzy clustering. This approach reduces the time required for the ant colony to find the optimal solution. Additionally, the clustering effect is enhanced by optimizing the ant colony algorithm within the sample distribution area identified in the first fuzzy clustering.

The steps to improve the ant colony algorithm are as follows:

Step 1: Set the maximum number of iterations to 0, initialize τ_{ij} and $\Delta\tau_{ij}$, m ants will be placed in the number of n vertices.

Step 2: Set the starting point of ants in the current solution set k according to probability ρ_{ij}^{k} move to next vertex j and put it in the current solution set. Among them, $k = 1, 2, ..., m$.

Step 3: Calculate the objective function value of each ant z_k and find the optimal solution. $k = 1, 2, ..., m$. The objective function formula is as follows in Eq. (1)

$$F = \sum_{k=1}^{i} d_k \tag{1}$$

where, d_k is the distance from the current cluster center to all the current sample points of the same kind, and the distance data is calculated using Euclidean distance.

Step 4: Modify the track strength based on the above steps, and follow the following Eq. (2).

$$\tau_{ij}(t+1) = \rho\tau_{ij}(t) + \Delta\tau_{ij} \tag{2}$$

where, $\rho \in (0,1)$, $\Delta\tau_{ij} = \sum_{k=1}^{m}\Delta\tau_{ij}^{k}$, $\Delta\tau_{ij}^{k}$ for No k only ants (i, j) pheromone concentration on the path, $\Delta\tau_{ij}$ is the pheromone concentration increment, ρ is the persistence of the current track.

Step 5: Update the number of iterations. If the current number of iterations does not reach the maximum number of iterations and all solutions found are different, return to Step 2. If the maximum number of iterations is reached or the same solution is found, the optimal solution under Eq. (1) is output, namely the cluster center.

Due to the noise in the actual industrial data, it is necessary to further process the labeled samples. Noise refers to other kinds of samples that appear in certain kinds of samples. The existence of noise can easily lead to fuzzy boundary of sample types and poor effect of subsequent classifiers. In view of this situation, this paper chooses to set the noise ratio α Analyze and screen each minority. The expression of noise ratio α is as follows.

$$\alpha = \frac{N_{\mathrm{M}}}{N_{\mathrm{N}}} \tag{3}$$

The K-nearest neighbor idea is introduced into the noise ratio discrimination process. N_{M} is the number of target samples in K neighborhood value, N_{N} is the number of samples of the non-target category of K neighboring value. Set the noise standard n, x is a sample set.

$$x_{\mathrm{M}} = \{x_{\mathrm{M1}}, x_{\mathrm{M2}}, x_{\mathrm{M3}}, x_{\mathrm{M4}}, x_{\mathrm{M5}}, x_{\mathrm{M6}}, \ldots\} \quad (4)$$

$$x_{\mathrm{N}} = \{x_{\mathrm{N1}}, x_{\mathrm{N2}}, x_{\mathrm{N3}}, x_{\mathrm{N4}}, x_{\mathrm{N5}}, x_{\mathrm{N6}}, \ldots\} \quad (5)$$

2.3 MCS-SVM

Traditional SVM is a widely used binary classification algorithm. Its basic model is the linear classifier with the largest gap in the feature space. Its core idea is to find the training set that can be divided according to the requirements, while maximizing the distance from the sample point to the hyperplane. The equation of hyperplane is as follows:

$$\omega x + b = 0 \quad (6)$$

For unknown samples with unknown tags, SVM cannot provide accurate classification results. However, when the sample distribution is unbalanced, the algorithm struggles to meet the requirements. Hence, this paper proposes an improved algorithm that combines K-Means with SVM. By leveraging the clustering effect of K-Means and the excellent classification effect of SVM, the issue of unfamiliar samples can be avoided. To address the problem of uneven sample distribution, new sample rules are introduced to enhance the distribution of the sample sets, thereby mitigating the impact of sample imbalance on calculation results. Additionally, to handle multi-classification samples more effectively, SVM is introduced multiple times in this paper. Iterative rules are set to expand the classifier, which initially deals with two-classification problems, to handle multi-classification problems.

When unbalanced samples are obtained, the first fuzzy classification of the samples is performed using SVM. It is assumed in this paper that the sample set contains at least two labels at this stage. If the obtained samples are completely unfamiliar, a K-Means clustering can be applied before the first fuzzy classification to obtain at least two labels. Based on the traditional SVM principle, the classification decision function of SVM is as follows:

$$f(x) = \mathrm{sign}\left(w^* \cdot x + b\right) \quad (7)$$

In the model, we will first follow the above process to conduct a fuzzy classification of sample points. Because the first classification is relatively fuzzy, and because the noise proportion coefficient is introduced in the subsequent calculation process to reduce the impact of noise, although the process may be affected by noise samples, the error caused by noise can be reduced by introducing relaxation variables.

In the aspect of error analysis, the accuracy of the final sample point based on the traditional algorithm under the original data and the accuracy of the improved algorithm based on the original data are calculated, and the advantages and disadvantages of the model in this paper are compared and analyzed.

According to the ACO-K-Means algorithm, more equipment status labels are obtained, which avoids the disadvantage that the traditional machine learning algorithm cannot quickly and effectively judge the equipment status in the face of fuzzy samples. At the same time, in order to have a better classification effect when the test set is input subsequently, it is necessary to carry out unbalance analysis on the clustered data. In order to ensure that the sample points of each state are relatively rich, this paper establishes a balanced proportion standard t at the same time, calculate the unbalance ratio of the current sample β.

$$\beta = \frac{X_{\max}}{X_{\mathrm{m}}} \quad (8)$$

where, $X_{\max}$ is the existing sample capacity under the label of the current target sample, X_{m} is the sample size of the current target sample. There β can set it freely according to actual needs. If $\alpha > \mathrm{N}$ and $\beta \geq t$ if it is satisfied at the same time, the following equation can be described as follows.

$$x_{\mathrm{new}} = x_i + \mathrm{rand}(0,1) \times (\bar{x} - x_i) \quad (9)$$

where, x_{new} is a new sample point, $\tilde{x}$ is the same sample point with the farthest Euclidean distance from the current cluster center, x_i is the cluster center of the sample point of the current category. This method can avoid the error caused by unbalanced data and noise data in the subsequent state classification, so that the model can deal with the fuzzy and unbalanced state label data in the actual industry.

Following the above method, after labeling the sample set, it is also necessary to introduce the SVM group to judge the equipment health status of the test set. At the same time, due to the actual industrial production, the health data of equipment often follows the time series and presents a certain growth trend, rather than disorder. Assuming that the improved K-Means algorithm obtains n health status, it needs to import $n - 1$ SVMs from a SVM group. This step is to overcome the disadvantage that traditional SVM cannot handle multi-classification problems.

2.4 Health Prognosis Process Based on ACO-K-Means and MCS-SVM

The health prognosis process based on ACO-K-Means and MCS-SVM algorithm is as follows.

Step 1: Input equipment data. There are at least two kinds of equipment health status. The experimental data is divided into training set and test set according to 2:1.

Step 2: SVM is used for the first fuzzy classification of the dataset, and the first classification result is obtained.

Step 3: Following the principle of K-Means algorithm, the first fuzzy clustering algorithm is used to compress the search area of the optimization algorithm, and then the ant colony algorithm and K-Means algorithm are used to cluster the classified sample points to obtain the qualified equipment status label.

Step 4: Introduce noise proportion coefficient n and balance proportion standard t at the same time, calculate the noise proportion coefficient of the current target sample α and unbalance ratio β. Add the clustered sample points according to the established rules.

Step 5: Use the data set of known tags to classify the sample points, and introduce the SVM set. If the number

of equipment status tags is n, the number of SVM is $n-1$. Complete the classification of sample points.

Step 6: Output the results, judge the equipment health status, fit the equipment health development trend and predict the future life of the equipment.

The model can avoid the impact of sample strangeness, sample imbalance, sample noise, etc. while maintaining high computing speed with small sample size, overcome the shortcomings of traditional algorithms and the disadvantages of the aforementioned data defects in the foreword scholars' research content, and provide theoretical basis for the actual prediction of enterprises.

3 CASE STUDY

3.1 Data Source

In this paper, the hydraulic pump of Caterpillar Company of America is used for simulation experiment. During the collection process, the health status of the hydraulic pump is mainly reflected by the bearing vibration data. The strain degree of the equipment can be calculated by observing the vibration frequency of the hydraulic pump bearing. The hydraulic pump was added with 20 - 80 mg of experimental materials, and vibration data were collected every 10 minutes. According to the characteristics of the data, the collected data were divided into four stages: bad, poor, medium and good. The sample point distribution of each state is determined by the improved K-Means algorithm, and the data index for evaluating each state is also determined by the corresponding SVM. There is no failure risk in the medium and good states, and there is failure risk in the bad and poor states.

3.2 Health Status Identification

Since there is no failure risk in the medium and good states, there is failure risk in the bad and poor phases. Tab. 1 lists the number of samples of hydraulic pump with and without failure risk. Due to the need of training SVM, about 2/3 of the data is used for training SVM, and about 1/3 of the data is used for testing.

Table 1 Hydraulic pump data set distribution table

Label		Risk of failure		No risk of failure	
Data allocation		14	7	12	7

Firstly, the collected data sets with and without fault risk are visualized. Due to much vibration data collected, this paper randomly selects vibration data in two directions for demonstration. The visualized data is shown in Fig. 2.

Blue is the sample distribution without failure risk, and red is the sample distribution with failure risk.

In actual industry, the data obtained often contains noise. In Fig. 2, it is obvious that there are several groups of noises. The influence of noise is ignored when SVM is introduced for the first time. Use SVM to solve the sample set for the first time, and the result is shown in Fig. 3.

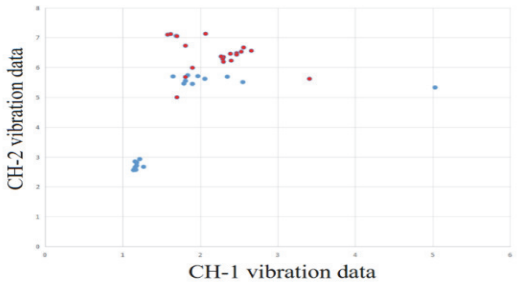

Figure 2 Vibration visualization of vibration data

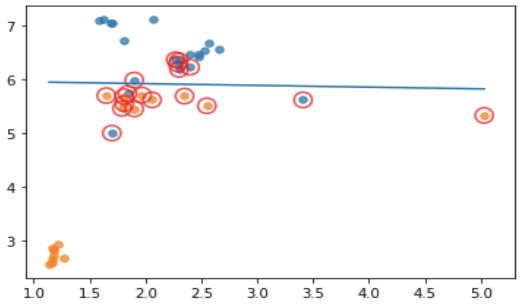

Figure 3 Display of the first SVM classification results

Where, the blue line is the calculated classifier, and the sample points circled in red circle are the support vector points. The distribution of sample points with and without fault risk can be obtained by using the first SVM classification. At the same time, there are noise samples in the figure due to the error of manual measurement or detection machine. The noise scale coefficient will be introduced in the subsequent experiments to process the noise sample points.

The improved K-Means algorithm is introduced based on the results of the first rough classification, and the clustering effect of the original algorithm is improved through the joint ant colony algorithm. According to the information, the health status of the equipment in this example shows four stages: bad, poor, medium and good. Therefore, we can know that the number of types of final output clustering results is 4. On this basis, the ant colony algorithm and K-Means algorithm are introduced to cluster the sample points after the first rough classification.

In order to reflect the advantages of the ant colony algorithm combined with K-Means algorithm, we first compare it with the traditional K-Means algorithm. Fig.4 shows the clustering effect of traditional K-Means algorithm:

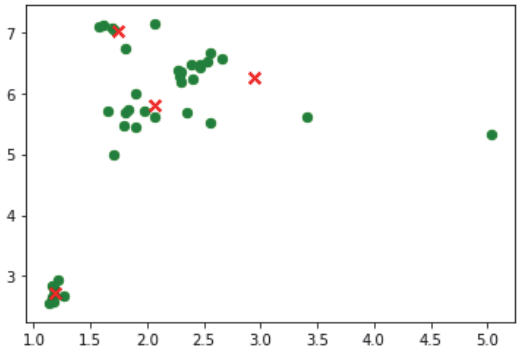

Figure 4 Cluster effect of traditional K-Means algorithm

Red "×" is the cluster center found by traditional K-Means.

Then use the K-Means algorithm optimized by the ant colony algorithm to cluster the data set. Firstly, the sample set is first fuzzy clustering to determine the search range of the ant colony. In order to ensure the optimization effect and shorten the search time of the ant colony, this paper compresses the search range of the ant colony to the maximum distribution area of each kind of sample points after the first fuzzy clustering, that is, the area divided between the extreme values of the coordinates of the sample points in the same cluster.

The extreme values of sample point distribution of health states can be obtained by calculation in Tab. 2.

Table 2 Extreme values of activity areas of equipment health status samples

Health status	Good	Medium	Poor	Bad
$x_{\min}$	1,14	1,65	1,58	2,27
$x_{\max}$	1,27	2,55	2,07	2,66
$y_{\min}$	2,57	5,45	5,00	6,19
$y_{\max}$	2,93	5,71	7,13	6,67

The ant colony algorithm is introduced into the search range according to the specified rules. In the K-Means algorithm, the maximum number of iterations is set to 100, and its initial value is set to 0. At the same time, initialization τ_{ij} and $\Delta\tau_{ij}$, 100 ants can be placed in each healthy state after the first fuzzy classification 4 extreme vertices. The cluster centers of the four health states of the searched devices are shown in Tab. 3.

Table 3 Display table of K-Means clustering center optimized by ant colony algorithm

Health status	Good	Medium	Poor	Bad
Horizontal coordinate of cluster center x	1,184	1,988	1,745	2,422
Vertical coordinate of cluster center y	2,73	5,597	7,032	6,415

To compare the clustering effect of the traditional K-Means algorithm and the K-Means algorithm optimized by the ant colony algorithm, this paper selects Dunn index as the evaluation index. Dunn equation is as follows:

$$d_{\min}\left(c_i, c_j\right) = \min_{x_i \in c_i,\, x_j \in c_j} \operatorname{dist}\left(x_i, x_j\right) \tag{10}$$

$$\operatorname{diam}(c) = \max_{1 \le i < j \le |c|} \operatorname{dist}\left(x_i,\, x_j\right) \tag{11}$$

$$DI = \min_{1 \le i \le k} \left\{ \min_{i \ne j} \left(\frac{d_{\min}\left(c_i, c_j\right)}{\max\limits_{1 \le l \le k} \operatorname{diam}\left(c_l\right)} \right) \right\} \tag{12}$$

The larger the Dunn index (DI), the better the clustering effect of the corresponding algorithm.

Through calculation, the Dunn index of the traditional K-Means algorithm and the ant colony algorithm proposed in this paper can be optimized and the average value of the Dunn index of the K-Means algorithm in each health state can be obtained. The numerical value and comparison results are shown in Tab. 4.

It can be seen in the table that the improved K-Means algorithm with the ant colony algorithm has better clustering effect while having high speed.

After observation, the sample point data in the bad state after ignoring the noise caused by human or detection machine shows an unbalanced state.

Table 4 Comparison of clustering effects between the traditional K-Means algorithm and the ant colony optimization K-Means algorithm in this paper

	Mean value of Dunn index	Clustering effect	Calculation speed
TraditionalK-Means	0,043	commonly	fast
Optimization of K-Means by Ant Colony Algorithm	0,067	good	fast

By evaluation, we can get that $K = 4$, and set $n = 0{,}1$. Follow the Eq. (3) to Eq. (5) to calculate the noise ratio coefficient of each sample point and compare it with 0,1.

Finally, three noise sample points were successfully selected, and the coordinate distribution of the sample points will be ignored in the subsequent calculation.

Calculate the unbalance ratio of the target sample at this time $\beta = 0{,}545$. Set this time $t = 0{,}8$; add the sample set according to Eq. (9) until $\beta \ge t$.

Then introduce the SVM set. At this time, the equipment health status presents four stages, and three SVMs are introduced. The final classification results are shown in Fig. 5 to Fig. 7.

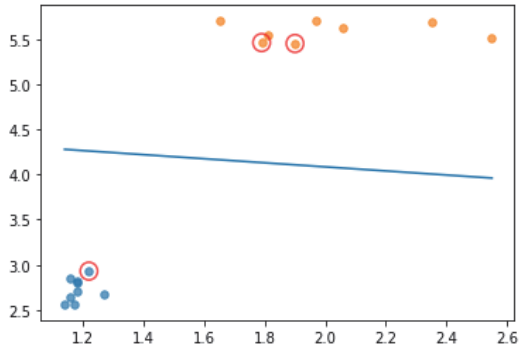

Figure 5 Classification diagram of good to medium status

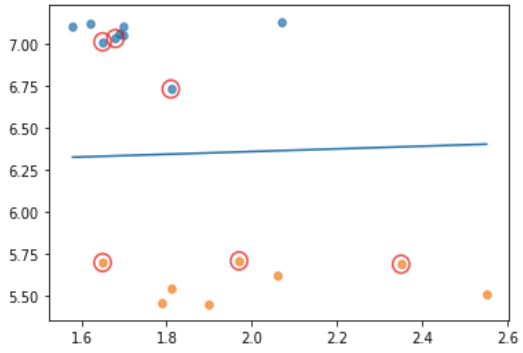

Figure 6 Classification diagram of medium to poor status

Under the simulation conditions in this paper, the data exhibit discrete trends and the sample values increase over time. Additionally, to address the issue of sample size, utilizing the SVM set in this study reduces computational time while maintaining classification accuracy.

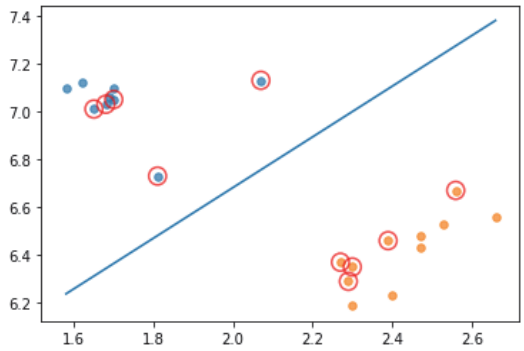

Figure 7 Classification diagram of poor to bad status

The comparison results between the ACO-K-Means combined MCS-SVM algorithm and the traditional machine learning algorithm proposed in this paper are shown in Tab. 5. Because the given original data set lacks the status label, the traditional K-Means is used to first label the sample set when comparing with some machine learning algorithms. When facing the two-dimensional classification problem, in addition to the SVM mentioned in this paper, KNN algorithm is one of the best algorithms in the field of machine learning when dealing with simple classification problems. Therefore, this paper focuses on the combination of traditional K-Means, ant colony optimization K-Means and SVM, KNN, multi-class self-adding SVM, and comparison of classification accuracy to prove the superiority of the proposed algorithm.

Table 5 Comparison of classification effects of various models

	Model	Identification accuracy
Traditional algorithm	K-Means-SVM	82,5%
	K-Means-KNN	77,5%
Joint algorithm	ACO-K-Means-SVM	83,7%
	ACO-K-Means-KNN	79,1%
	ACO-K-Means-MCS-SVM	93%

Clustering algorithms are employed to label them before utilizing traditional machine learning algorithms for calculations. K-Means and improved K-Means are used here for clustering to control variables and facilitate better comparisons. Results indicate that the recognition accuracy of traditional machine learning algorithm combinations is lower compared to that of ACO-K-Means combined with corresponding classification algorithms. The comparative analysis demonstrates that the proposed combination of ACO-K-Means and MCS-SVM exhibits superior recognition accuracy. The addition of new sample points helps mitigate errors resulting from sample imbalance and overfitting due to limited samples.

Regarding cost and algorithm complexity analysis, the proposed model in this paper demonstrates comparable calculation speeds to other algorithms listed in Tab. 5, while achieving higher accuracy and operating on relatively small datasets. Thus, the proposed model outperforms in terms of both cost analysis and algorithm complexity analysis.

3.3 Health Prognosis Results

The health status of the hydraulic pump is mainly reflected by the calculated RMS value (root mean square value of vibration) under its bearing vibration data. The RMS value can be calculated and analyzed to quickly determine the health status of the equipment and predict the remaining life of the equipment, providing a reference basis for enterprises to continue to use the hydraulic pump. The calculation formula of RMS is as follows.

$$X_{\text{rms}} = \sqrt{\frac{X_1^2 + X_2^2 + \ldots + X_N^2}{N}} \tag{13}$$

Fig. 8 shows the fitted RMS change trend of the hydraulic pump.

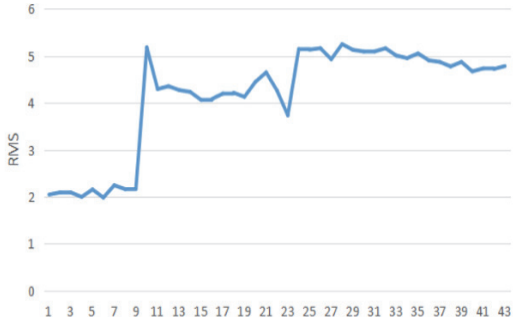

Figure 8 Change trend of hydraulic pump RMS

The abscissa represents the test time of the equipment, and the ordinate represents the RMS value. When the equipment's RMS value approaches 5, the corresponding time point for the equipment to enter the health decline period is 25. At this time, the equipment is at risk of failure. However, at time point 10, an outlier is observed in the RMS distribution, which does not align with the actual industry's equipment operation trend. Therefore, the data at this point should be disregarded when using the RMS distribution chart to predict the service life of the hydraulic pump. The subsequent analysis focuses on data points with RMS values greater than 5 to predict the remaining equipment life at this stage. Data points with RMS values less than 5 will not be considered.

Table 6 RMS and time distribution of equipment under failure risk

Test time / min	Hydraulic pump bearing RMS
250	5,143
260	5,133
270	5,163
280	4,928
290	5,250
300	5,128
310	5,092
320	5,094
330	5,162
340	5,006
350	4,952
360	5,052
370	4,904
380	4,871
390	4,776
400	4,871
410	4,670
420	4,733
430	4,721
440	4,782

The relevant equipment data for prediction has been filtered in the following table.

Use the data in the table to calculate the remaining life of the equipment. The maximum service time of the equipment is the 440th minute of the test time. The remaining life of the equipment can be calculated by calculating the difference between the maximum service time of the test and the current service time of the test. Analyze the calculated residual life and corresponding data points and fit the residual life curve of the equipment. First, use ACO-K-Means-MCS-SVM to fit the change trend of the RMS of the device.

The comparison results between the fitting curve and the real value are shown in Fig. 9.

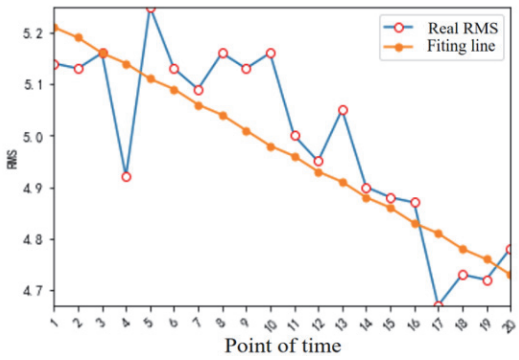

Figure 9 Comparison between real RMS and fitting line of equipment

By comparing the real RMS of the equipment with the trend line fitted, it can be concluded that the model proposed in this paper can accurately predict the future RMS of the equipment. There are some abnormal values in the figure due to the error of manual or detection machine. In actual industry, it is impossible to successfully eliminate all noise, so in order to fit the authenticity of the curve, some abnormal values are selected to be taken into account in the simulation process.

Second, the future trend of the equipment is predicted by considering RMS and detection time, shown in Fig. 10.

The comparison results show that the predicted RUL after data point 10 completely coincides with the true value. Due to the rule of adding new sample points in ACO-K-Means-MCS-SVM, the RUL prediction value before data point 10 has certain error. However, with the progress of detection, the error will gradually decrease to disappear.

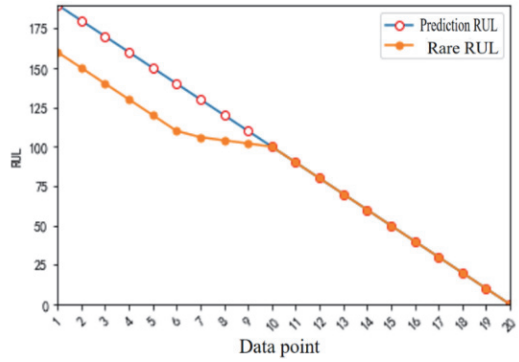

Figure 10 Comparison between predicted RUL and true value of equipment

According to the above prediction results of RMS and RUL, the model proposed in this paper is still highly effective in life prediction. The model proposed in this paper has high practicability and can meet the requirements of enterprises for the accuracy of equipment condition analysis and life prediction in the case of small samples, unbalanced samples and fuzzy sample labels that are likely to occur in the actual industry. At the same time, the remaining service life of the equipment can provide the basis for the enterprise to replace the equipment in advance and avoid the economic losses caused by the equipment life judgment error in actual production.

4 CONCLUSION

In this paper, samples are classified into 'with failure risk' and 'without failure risk' using SVM fuzzy classification. An improved K-Means algorithm is then employed to label the fuzzy dataset for determining equipment health status. The ant colony algorithm is introduced to enhance the K-Means algorithm by reducing the search range through fuzzy clustering. It efficiently finds cluster centers meeting the conditions. To achieve multi-classification and mitigate errors caused by sample imbalance and noise, this paper introduces noise proportion coefficient, noise neglect rule, sample imbalance proportion coefficient, and a new rule. These rules prevent errors from noise points and unbalanced samples. Subsequently, the SVM set is utilized to judge the health status of future equipment health data. ACO-K-Means demonstrates superior clustering while MCS-SVM exhibits better classification accuracy. Moreover, the combined ACO-K-Means and MCS-SVM model effectively predicts the future development trend of equipment lifetime and provides a reference for equipment replacement. The proposed model performs well in datasets with small sample imbalance and fuzzy labels, while mitigating interference from noisy data.

This paper employs multiple models and algorithms to establish suitable models for handling noise and imbalanced samples, prioritizing efficiency and minimal resource consumption, resulting in high experimental accuracy. However, the performance of the proposed model is only validated through examples, lacking comprehensive evaluation across other datasets and metrics. Future research will focus on refining parameter selection and tuning methods, conducting extensive experimental validation, and testing the model's robustness with additional datasets. These efforts aim to further enhance the model's performance and interpretability.

Acknowledgements

The work presented in this paper has been supported by grants from the National Natural Science Foundation of China (Nos. 72271161), Action Plan for Scientific and Technological Innovation of Shanghai Science and Technology Commission (No.21SQBS01404). The authors are indebted to the reviewers and editors for their constructive comments, which greatly improved the contents and exposition of this paper.

5 REFERENCES

[1] Behera, S., Misra, R., & Sillitti, A. (2021). Multiscale deep bidirectional gated recurrent neural networks based

prognostic method for complex non-linear degradation systems. *Information Sciences*, *554*, 120-144. https://doi.org/10.1016/j.ins.2020.12.032

[2] Dong, M. & He, D. (2007). Hidden semi-Markov model-based methodology for multi-sensor equipment health diagnosis and prognosis. *European Journal of Operational Research*, *178*(3), 858-878. https://doi.org/10.1016/j.ejor.2006.01.041

[3] Duan, C., Makis, V., & Deng, C. (2019). Optimal Bayesian early fault detection for CNC equipment using hidden semi-Markov process. *Mechanical Systems & Signal Processing*, *122*, 290-306. https://doi.org/10.1016/j.ymssp.2018.11.040

[4] Garrido, J., Yu, W., & Li, X. (2016). Robot trajectory generation using modified hidden Markov model and Lloyd's algorithm in joint space. *Engineering Applications of Artificial Intelligence*, *53*, 32-40. https://doi.org/10.1016/j.engappai.2016.03.006

[5] Guha, A., Patra, A., & Vaisakh, K.V. (2017). Remaining useful life estimation of lithiumion batteries based on the internal resistance growth model. *Control Conference IEEE*, 33-38. https://doi.org/10.1109/INDIANCC.2017.7846448

[6] Huang, L., Huang, S., & Lai, Z. (2020). On the optimization of site investigation programs using centroidal Voronoi tessellation and random field theory. *Computers and Geotechnics*, *118*, 103331. https://doi.org/10.1016/j.compgeo.2019.103331

[7] Huang, W. & Dietrich, D. L. (2005). An alternative degradation reliability modeling approach using maximum likelihood estimation. *IEEE Transactions on Reliability*, *54*(2), 310-317. https://doi.org/10.1109/TR.2005.845965

[8] Kwon, J. (2020). Particle swarm optimization-Markov Chain Monte Carlo for accurate visual tracking with adaptive template update. *Applied Soft Computing*, *97*, 105443. https://doi.org/10.1016/j.asoc.2019.04.014

[9] McLeay, T., Turner, M. S., & Worden, K. (2021). A novel approach to machining process fault detection using unsupervised learning. Proceedings of the Institution of Mechanical Engineers, Part B: *Journal of Engineering Manufacture*, *235*(10), 095440542093755. https://doi.org/10.1177/0954405420937556

[10] Li, J., Zhang, X., Zhou, X., & L, Lu. (2019). Reliability assessment of wind turbine bearing based on the degradation-Hidden-Markov model. *Renewable Energy*, *132*, 1076-1087. https://doi.org/10.1016/j.renene.2018.08.048

[11] Liu, J., Li, Q., Chen, W., & Cao, T. (2018). A discrete hidden Markov model fault diagnosis strategy based on K-means clustering dedicated to PEM fuel cell systems of tramways. *International Journal of Hydrogen Energy*, *43*(27), 12428-12441. https://doi.org/10.1016/j.ijhydene.2018.04.163

[12] Liu, Q., Dong, M., Lv, W., Geng, X., & Li, Y., (2015). A novel method using adaptive hidden semi-Markov model for multi-sensor monitoring equipment health prognosis. *Mechanical Systems & Signal Processing*, *64-65*, 217-232. https://doi.org/10.1016/j.ymssp.2015.03.029

[13] Liu, Q., Dong, M., & Peng, Y. (2012). A novel method for online health prognosis of equipment based on hidden semi-Markov model using sequential Monte Carlo methods. *Mechanical Systems & Signal Processing*, *32*, 331-348. https://doi.org/10.1016/j.ymssp.2012.05.004

[14] Lv, X., Wang, H., Zhang, X., Liu, Y., Jiang, D., & Wei, B. (2021). An evolutional SVM method based on incremental algorithm and simulated indicator diagrams for fault diagnosis in sucker rod pumping systems. *Journal of Petroleum Science and Engineering*, *203*, 108806. https://doi.org/10.1016/j.petrol.2021.108806

[15] Li, X. Q., Jiang, H., & Liu, S. (2021). A unified framework incorporating predictive generative denoising autoencoder and deep Coral network for rolling bearing fault diagnosis with unbalanced data. *Measurement*, *178*. https://doi.org/10.1016/j.measurement.2021.109345

[16] Peng, Y. & Dong, M. (2011). A prognosis method using age-dependent hidden semi-Markov model for equipment health prediction. *Mechanical Systems & Signal Processing*, *25*, 237-252. https://doi.org/10.1016/j.ymssp.2010.04.002

[17] Wu, B., Li, W., & Qiu, M. Q. (2017). Remaining useful life prediction of bearing with vibration signals based on a novel indicator. *Shock and Vibration*, 8927937. https://doi.org/10.1155/2017/8927937

[18] Yan, Y., Cai, J., Li, T., Zhang, W., & Sun, L. (2021). Fault prognosis of HVAC air handling unit and its components using hidden-semi Markov model and statistical process control. *Energy and Buildings*, *240*, 110875. https://doi.org/10.1016/j.enbuild.2021.110875

[19] Bhouri, M. A. & Perdikaris, P. (2021). Gaussian processes meet Neural ODEs: a Bayesian framework for learning the dynamics of partially observed systems from scarce and noisy data. *Philosophical Transactions of the Royal Society A*, *380*(2229), 20210201-20210201. https://doi.org/10.1098/rsta.2021.0201

[20] Maddu, S., Cheeseman, B. L., Sbalzarini, I. F., & Müller, C. L. (2022). Stability selection enables robust learning of differential equations from limited noisy data. *Proceedings of the Royal Society A*, *478*(2262*)*, 20210916-20210916. https://doi.org/10.1098/rspa.2021.0916

Contact information:

Qinming LIU, PhD, Professor
Department of Industrial Engineering, Business School,
University of Shanghai for Science and Technology,
516 Jungong Road, Shanghai, 200093, P. R. China
E-mail: qmliu@usst.edu.cn

Fengze YUN, Postgraduate,Student
Department of Industrial Engineering, Business School,
University of Shanghai for Science and Technology,
516 Jungong Road, Shanghai, 200093, P.R. China
E-mail: Rainlinwind@foxmail.com

Ming DONG, PhD, Professor
Department of Operations Management, Antai College of Economics & Management, Shanghai Jiao Tong University,
1954 Huashan Road, Shanghai, 200030, P.R. China
E-mail: mdong@sjtu.edu.cn

Darko DJORIC, MSc, CEO
MIND Group,
Aleja Milanović bb,34325 Kragujevac, Serbia
E-mail: darko@djoric.rs

Nikola ZIVLAK, PhD, Associate Professor
(Corresponding author)
Department of Industrial Engineering and Management, Faculty of Technical Sciences,
University of Novi Sad,Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia
E-mail: nikola.zivlak@uns.ac.rs