J. Shanghai Jiao Tong Univ. (Sci.), 2019
https://doi.org/10.1007/s12204-019-2107-1

Feature Recognition and Selection Method of the Equipment State Based on Improved Mahalanobis-Taguchi System

WANG Ning[1*] (王　宁), *ZHANG Zhuo*[2] (张　卓)
(1. School of Automobile, Chang'an University, Xi'an 710064, China;
2. School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Mahalanobis-Taguchi system (MTS) is a kind of data mining and pattern recognition method which can identify the attribute characteristics of multidimensional data by constructing Mahalanobis distance (MD) measurement scale. In this paper, considering the influence of irregular distribution of the sample data and abnormal variation of the normal data on accuracy of MTS, a feature recognition and selection model of the equipment state based on the improved MTS is proposed, and two aspects of the model namely construction of the original Mahalanobis space (MS) and determination of the threshold are studied. Firstly, the original training sample space is statistically controlled by the X-bar-S control chart, and extreme data of the single characteristic attribute is filtered to reduce the impact of extreme condition on the accuracy of the model, so as to construct a more robust MS. Furthermore, the box plot method is used to determine the threshold of the model. And the stability of the model and the tolerance to the extreme condition are improved by leaving sufficient range of the variation for the extreme condition which is identified as in the normal range. Finally, the improved model is compared with the traditional one based on the unimproved MTS by using the data from the literature. The result shows that compared with the traditional model, the accuracy and sensitivity of the improved model for state identification can be greatly enhanced.
Key words: Mahalanobis-Taguchi system (MTS), extreme condition, X-bar-S control chart, box plot method, Mahalanobis space (MS), Mahalanobis distance (MD), threshold, feature recognition, equipment state
CLC number: N 941.4 **Document code:** A

0 Introduction

Mahalanobis-Taguchi system (MTS)[1-2] is a highly robust and adaptable statistical method applying statistical prediction theory and feature selection theory to the engineering of quality. The basic idea of the MTS is to define a measurement table and construct a Mahalanobis space (MS) based on the collected normal training samples that makes up the benchmark space. And then, the benchmark space is used to analyze and judge the new test sample according to its range. If the newly collected sample is within the range of the benchmark space, it is considered to be same as the previously collected normal training sample. And then the Mahalanobis distance (MD) is used to measure the space range of MTS. In other words, if the MD of the test sample is less than the threshold that determined by the training sample set, the test sample is regarded as falling into the benchmark space, that is to say, it is normal. On the other hand, if the test sample falls outside the benchmark space, we can judge that the larger the MD of the test sample is, the farther the distance between the test sample and the boundary of the benchmark space. Finally, orthogonal array (OA) and signal-noise ratio (SNR) are used in the MTS to select the valid attribute characteristics from the constructed benchmark space. Through the above-mentioned methods, the selected valid variable is used to define the new measurement table, which is more effective than the old one.

MTS can provide powerful support for feature recognition and dimension reduction of the multidimensional and complex system, and produce great application effect and economic benefit. Recently, MTS has made great progress in theoretical researches and industrial applications, especially in pattern recognition[3-4], manufacturing process diagnosis[5], disease diagnosis and prediction[6], product monitoring and testing[7-8]. At

Received date: 2018-07-31
Foundation item: the National Natural Science Foundation of China (No. 71401016), the Shaanxi Provincial Natural Science Foundation of China (No. 2019JM-495), and the Fundamental Research Funds for Central Universities of Chang'an University (Nos. 300102228110 and 300102228402)
*E-mail: ningwang@chd.edu.cn

the same time, other researches related to MTS, such as the determination of threshold, the improvement of measurement scale of the model[9] and the construction of the MS, were deeply studied.

According to the above discussion, the determination of threshold for MTS is critical to enhance the ability of the model to identify the equipment state and the robustness of the judgment. In terms of the researches on the determination of threshold, initially, Taguchi and Jugulum[1] determined the threshold by constructing a mass loss function that relies on the misjudgment loss minimization criterion of the normal and abnormal samples. However, the loss value in the function depends on the subjective experience of the professionals. Nakatsugawa and Ohuchi[10] used Gamma distribution to describe the cumulative distribution of the training samples and determine the threshold. Then, the confidence interval of the threshold determined by the method was calculated by the Kolmogorov-Smirnov test. But there is lack of enough rigorous theoretical proof to prove the method. Furthermore, the Kolmonov-Smirnov test is always used for the test of one-dimensional distribution, which has great limitations in the practice for the multivariate distribution. Su and Hsiao[11] proposed a probabilistic threshold model, which used the Chebyshev's inequality to make the misjudgment rate of the normal sample lowest. However, the model does not take into account the misjudgment of the abnormal sample. Xu et al.[12] assumed that the MD of the training sample was subject to the F-distribution and the threshold was determined based on the number of variables. This method also lacks of rigorous theoretical proof due to the irregular distribution of actual samples. In addition, from the perspective of the distribution or algorithms, many other methods were proposed to determine the threshold of the model[13-15]. Most of the methods have poor stability, and take longer calculation time. Meanwhile, most of them didn't consider the irregular distribution of actual data and the influence of initial MS on the threshold and the accuracy of the model. Therefore, the fault tolerance of the model to the extreme condition of sample data is low, and the failure of feature recognition will occur while the data are fluctuating.

On the construction of initial MS, the existing research is only carried out from the perspective of screening the "normal" from the normal training sample, but does not consider the "abnormal situation" in the normal sample to be the extreme condition in the normal data range, thus ignores the impact of the construction of initial MS on the threshold. For example, Wang et al.[16] argued that the extreme values in the training sample data that would be used to construct the MS were abnormal, and they should be accepted as abnormal sample adding to the test sample set. But this method ignored the relationship between the MS construction and the threshold determination, which will lead to the results in the original space with a very poor effect. Liparas et al.[17-18] used the two-step clustering to similarly screen out the abnormal values of the training samples, but ignored the variation of the threshold and the fluctuation of the normal sample. And the two-step clustering method is limited in practical application, while the actual data do not necessarily obey the normal distribution, and the linear relationship among the characteristic variables is difficult to determine.

On the basis of the above-mentioned researches, this paper presents a feature recognition and selection model of equipment state based on the improved MTS that improves the construction of initial MS and the threshold determination. Firstly, the original training sample is controlled by X-bar-S control chart to filter the extreme values in the normal range which has great impact on the MS, so that the benchmark space can be more accurate to reflect the characteristics of the normal state of the equipment. Then, the box plot method is used to analyze the MD calculated by the normal training sample set to determine the threshold. And this method will eliminate the interference of the irregular distribution and the fluctuation of data. Finally, the improved model is compared with the traditional MTS-based state feature recognition and selection model for the equipment. And the result shows that the improved model in terms of the accuracy and sensitivity for recognizing the equipment state has a great improvement, and it also shows that the method of considering the relationship between the threshold determination and the construction of MS is more accurate and stable than that without considering the relationship.

1 Traditonal Model

1.1 Construction of MS

According to the actual application and obtained sample data, the required state features are obtained from different dimensions of the normal state of the equipment, and then a measurement table is defined. Then the MS is constructed on the basis of the data from the normal sample and the MD of the normal sample is calculated by using the inverse matrix method[1-2].

The calculation process is as follows:

Step 1 Calculate the mean of the state feature for the normal sample:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}. \tag{1}$$

Step 2 Calculate the standard deviation of the feature for the normal:

$$s_j = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}. \tag{2}$$

Step 3 Normalize the data of each feature, obtain the normalized matrix of the original data ($\boldsymbol{Z}_{n\times m}$), and then transpose it. The element can be calculated by

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}. \tag{3}$$

Step 4 Calculate the correlation matrix ($\boldsymbol{C}_{m\times m}$):

$$\boldsymbol{C}_{m\times m} = \frac{1}{n-1}\boldsymbol{Z}^{\mathrm{T}}_{n\times m}\boldsymbol{Z}_{n\times m}. \tag{4}$$

Step 5 Calculate the inverse matrix ($\boldsymbol{C}^{-1}_{m\times m}$) of the correlation matrix ($\boldsymbol{C}_{m\times m}$).

Step 6 Calculate the MD for the normal sample, which constitutes the benchmark space:

$$\mathbf{MD} = [\mathrm{MD}_1 \ \ \mathrm{MD}_2 \ \ \cdots \ \ \mathrm{MD}_n] = \frac{1}{k}\boldsymbol{Z}_{n\times m}\boldsymbol{C}^{-1}_{m\times m}\boldsymbol{Z}^{\mathrm{T}}_{n\times m}, \tag{5}$$

where, n is the number of samples of the normal running state for the equipment; m is the number of the original state feature variable of the equipment; x_{ij} is the j-th feature in the i-th observation of the training sample of the state for the equipment, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$; $\bar{x}_j$ is the mean of the j-th feature in the normal training sample of the state for the equipment; s_j is the standard deviation of the j-th feature in the normal training sample of the state for the equipment; **MD** is the Mahalanobis distance of the training sample of the state for the equipment; k is the number of the state feature in the original measurement table, $1 \leqslant k \leqslant n$.

1.2 Validity Verification of Benchmark Space

Firstly, the abnormal condition of the equipment running state should be determined according to practical application. And in the practical application, the abnormal condition is defined based on the data of different deterioration degrees of the equipment state. In the paper, in order to verify the validity of the measurement table and the benchmark space, the abnormal condition of the equipment is defined as all states except the normal sample space.

After determining the abnormal condition of the equipment running state, the MD calculation of the observation under the abnormal condition will be performed. And it is worth noting that the MD calculation for the abnormal condition is on the basis of the MS constructed by the normal sample, that is to say, we normalize the data of the abnormal observation and calculate the MD for the abnormal observation by using the eigenvalues (the mean, the standard deviation and the correlation matrix) that were calculated by the normal training sample (seen Eqs. (3) and (5)).

According to the basic view of the MTS, the benchmark space which is constructed according to the normal sample should have a certain range. If the MD of the observation falls within this range, the value of the MD for the sample should be near or less than a certain threshold. And if the observation is not within the benchmark space, the value of the MD will increase as the difference between the observation and the normal sample set constructing the MS. In this paper, we use the inverse matrix method to calculate the MD. And according to the Ref. [1], the MD of the benchmark space calculated by this method should be distributed at about 1, or at least not more than a threshold near 1. That is to say, the threshold can be determined by the mean of the MD from the benchmark space or 1 to verify the validity of the benchmark space[21]. Based on this discussion, the different values of the MD correspond to the different state characteristics of the equipment, and we can compare the MD calculated by the sample data in the normal state and the MD of the sample data in the abnormal condition to verify the validity of the benchmark space.

1.3 Recognization and Selection of State Feature

The running state of the equipment is collectively characterized by different state features. However, not all state features have a positive effect on characterizing the running states of the equipment. Therefore, the MTS uses OA to generate different combinations of the state features to build different measurement tables. And in general, the choice of the OA depends on the number of state feature variables. In the MTS, the OA containing two levels is used, and the level of 1 to select the corresponding state feature variables to constitute a new measurement table, level of 2 indicates that the feature variable is not selected. So each row in the OA represents a new measure table constructed by the feature variable of level 1.

Then after calculating the MD of the different measurement tables, the larger-the-better type SNR will be used to assess the robustness for the different measurement tables. The SNR can be calculated as

$$\eta_q = -10\lg\left(\frac{1}{t}\sum_{r=1}^{t}\frac{1}{\mathrm{MD}_r}\right), \tag{6}$$

where, η_q is the SNR of the combination of the q-th line feature variable in the OA, and in the paper we use the OA of two levels ($L_{12}(2^{11})$), $1 \leqslant q \leqslant 12$; t is the number of the abnormal samples; MD_r stands for the MD of the r-th abnormal observation, and $1 \leqslant r \leqslant t$;

And then the mean of SNR is calculated for each feature variable under different levels of each feature to determine the different recognition effects between the presence of the state feature and the absence of the state feature. And then we can determine whether the state feature variable is valid to recognize the equipment running state. The gain of SNR is shown as

$$\mathrm{Gain}(i) = \mathrm{Mean}(i)_{\text{level-1}} - \mathrm{Mean}(i)_{\text{level-2}}, \tag{7}$$

where, Gain(i) is the gain of SNR for the i-th feature variable; Mean$(i)_{\text{level-1}}$ is the mean of SNR of the i-th feature variable under level 1; Mean$(i)_{\text{level-2}}$ is the mean of SNR of the i-th feature variable under level 2. If the gain of a feature variable is greater than 0, the feature variable is selected to form a new measurement table, otherwise, the feature variable is removed from the table.

2 Model Based on Improved MTS

2.1 Construction of MS Based on X-Bar-S Control Chart

The X-bar-S control chart is a method in the statistical process control (SPC), which is used for quality engineering to detect, monitor, evaluate, and report the abnormalities in the quality process. In the X-bar-S control chart, a center line indicates the mean of the process variable which is controlled by the chart, and is on the upper and lower sides of the center line. And there are also two lines to control the boundary, which are respectively called as the upper control limit (UCL) and the lower control limit (LCL) both determined by the standard deviation of the process variable. With the process variable changing over time, if the observation is between the UCL and the LCL, the state of the process is considered to be normal; otherwise, if the observation falls outside the limit, the process state is abnormal and the process is out of control.

Since the running state of the equipment is collectively characterized by numerous different state features, the process "abnormality" of the single state feature variable cannot certainly determine the running state of the equipment. And after analyzing the normal historical data of the multidimensional state feature for the equipment, a part of the state feature data should be out of the control limit boundary. So for the normal state of the equipment, we can just only consider that these data which fall out of the boundary are the extreme change of the normal range, which indicates these data are not abnormal. But for the construction of the MS, the feature value of the MS including the mean, the standard deviation and the correlation matrix will be affected by the extreme changes of these data, which will affect the accuracy of the model and lead to wrong judgement.

So based on this, this paper proposes a new method of constructing the MS on the basis of the X-bar-S control chart, which is used to filter the original data and remove the effect of the extreme change of the state feature on the model. And the method will cause the model based on the training sample data to be more concentrated and accurate. And for accuracy comparison of the method, the $\pm 3\delta$ and the $\pm 2\delta$[21] are used respectively as the control limit of the chart to build two different models which have different degrees of control.

2.2 Threshold Determination Method Based on Box Plot Method

From the previous discussion, we can see that while choosing the threshold we should not only consider the ability of the threshold to recognize the normal state of the equipment, but also consider the ability of the threshold to recognize the abnormal state. But the MD of the normal sample data and the abnormal sample data has a certain range of overlap, which will make it difficult to determine a decided threshold to recognize the equipment state. In this paper, a determination method of the threshold is proposed on the basis of the box plot method based on the MS constructed by X-bar-S control chart. Firstly, the box plot is used to describe the MD calculated by the benchmark space, and the condition of the MD distribution of the normal state which has filtered the extreme changes will be shown in the result.

In detail, the box plot method is to use five statistics among the data: the minimum, the first quartile (Q_1), the median, the third quartile (Q_2) and the maximum to roughly describe the condition of the data. We can get a rough summary of the symmetry, the discreteness and the skew of the data through the box plot, and then combined with the interquartile range (IQR), we can determine the upper and lower limits of the box plot for the data. And generally, the box plot can be used to identify abnormal values and extreme values of the data sets[24].

In the MS constructed by the X-bar-S control chart, the distribution of the data is more regular and centralized, and the MD based on the data will truthfully reflect the state features of the equipment. Based on that, the "outlier and extreme values" identified by the box plot should not be treated as an abnormal condition, but be accepted as a "boundary state" which is minority but normal for the running state of the equipment. In conjunction with the discussion of the MTS criterion in Section 1, under the MS based on the control chart, we use the third quartile as the threshold to exclude double interference of the extreme condition of the original data and the extreme condition of the MD to the accuracy of model. Because the quartile has a strong anti-jamming capability and is more insensitive to the extreme values, in which as many as 25% of the data are allowed to have significant changes and perturbations not affecting the stability of the model and its threshold.

Therefore, on the basis of the X-bar-S control chart and the box plot method, the feature recognition and selection method of the equipment state based on the improved MTS is shown in Fig. 1.

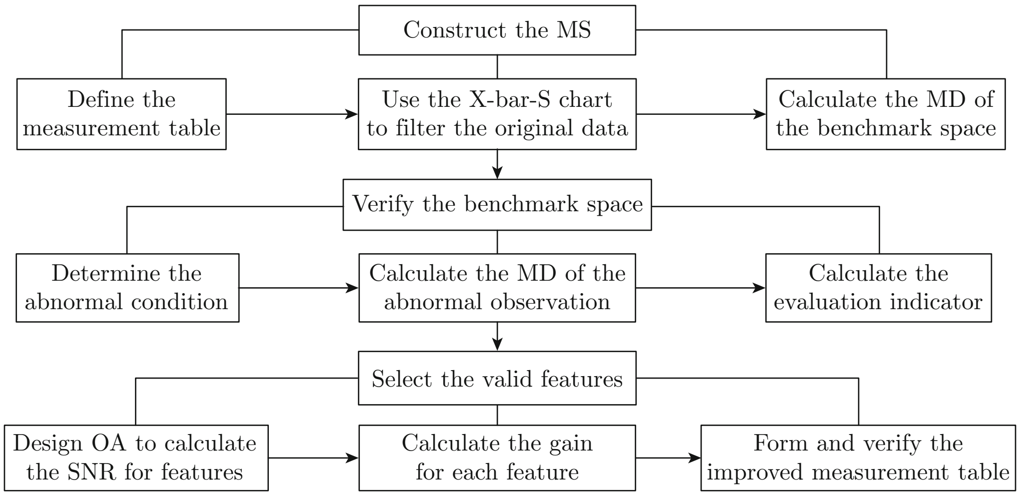


Fig. 1 The model based on the improved MTS

3 Model Robustness Evaluation

When we use different models to identify the normal state and abnormal state of the equipment after determining different measurement tables, there are often four different results as follows[16]:

(1) True positives (TP): the abnormal samples which are correctly recognized to be abnormal by the model.

(2) False negatives (FN): the abnormal samples which are falsely recognized to be normal by the model (Type II error).

(3) True negatives (TN): the normal samples which are correctly recognized as the normal by the model.

(4) False positives (FP): the normal samples which are falsely recognized to be abnormal by the model (Type I error).

In order to evaluate the accuracy of the judgments and the effectiveness of the model, there are four indicators including the Accuracy, the Sensitivity, the Specificity and the Relativity Sensitivity (RS) to evaluate the model. And the equations are

$$\text{Accuracy} = (\text{TN} + \text{TP})/(\text{TN} + \text{TP} + \text{FN} + \text{FP}), \quad (8)$$

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}), \quad (9)$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}), \quad (10)$$

$$\text{RS} = \text{Sensitivity}/\text{Specificity}. \quad (11)$$

From the calculation of the indicators, it can be seen that the Accuracy indicates the classification ability of the model to the equipment state, the Sensitivity characterizes the ability of the model to determine the abnormal state of the equipment, the Specificity characterizes the ability of the model to determine the abnormal state of the equipment, the RS characterizes the balanced ability of the model to determine the normal state and the abnormal state of the equipment[16].

4 Experimental Verification

In this section, we will use the rolling bearing feature recognition experiment in which the initial data are from Ref. [21] as the verification experiment to analyze and compare the validity and the accuracy for the improved model and the traditional one.

4.1 Data Collection and Benchmark Space Construction

Firstly, we extract eleven state feature variables of the bearing running state to form the initial measurement table from the state feature recognition experiment of the rolling bearing. And these features are the failure frequency of the bearing guard, the frequency of the rolling body through the inner ring, the frequency of the rolling body through the outer ring, the root mean square of the vibration signal, the kurtosis of the vibration signal and the temperature. There are 1 500 sets of the normal state data and 3 839 sets of the abnormal state data.

Then for comparison, we screened the sample data based on the X-bar-S control chart by respectively constructing the control limit of $\pm 3\delta$ and $\pm 2\delta$, and there were 1 289 normal samples in the control limit of $\pm 3\delta$ and 1 182 normal samples in the control limit of $\pm 2\delta$. Then the samples were randomly divided into training samples and test samples according to the ratio of 4 : 1, and the results are shown in Table 1.

Next in the packet of Table 1, we used 1 192 training normal samples to construct the traditional MS (MS_0) according to the equations Eqs. (1)—(4), and similarly we used the packets data of 1 027 training normal samples and 951 training normal samples in the $\pm 3\delta$ and $\pm 2\delta$ control limit to respectively construct MS_1 and MS_2 which have different degrees of control by the control chart. And the rest sample data were used to verify the validity and accuracy of the model.

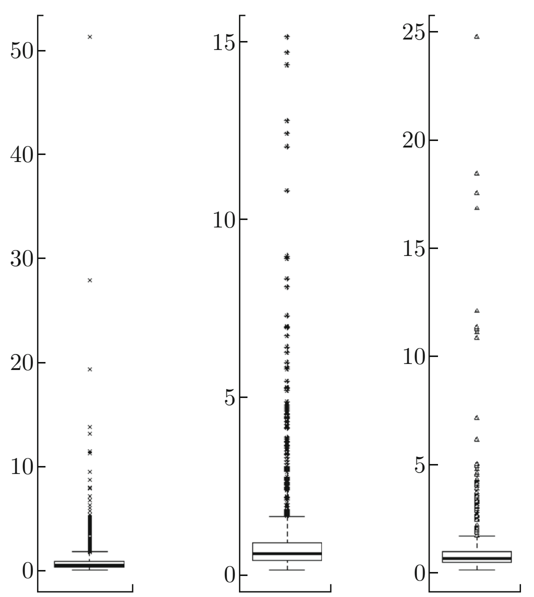
Table 1 Construction of the MS

Data	Total Samples		Training Sample		Test Sample	
	Normal	Abnormal	Normal	Abnormal	Normal	Abnormal
Original	1 500	3 839	1 192	3 062	308	777
$\pm 3\delta$	1 289	3 363	1 027	2 661	262	702
$\pm 2\delta$	1 182	2 988	951	2 360	231	628

4.2 Threshold Determination

The MDs of three different benchmark spaces were calculated according to Eq. (5). According to the box plot method, the box plots of three sets of MD data were used to determine the thresholds of the three models.

In Fig. 2, we can see that there are some extreme values in the three benchmark spaces, and for comparison, the distribution tends of the data of the two spaces constructed by the control chart are more concentrated than the data of the original space. In the benchmark spaces constructed by the control chart, there are more data falling between the first quantile and the third quantile than the original space. And the size of the extreme values is much less than that in the original space, which indicates that the spaces based on the control chart are better to be used to characterize the distribution of the normal state data. And compared with the space constructed by the $\pm 3\delta$ control limit, the filtering of the data in the benchmark space based on the $\pm 2\delta$ control limit is somewhat excessive to lead to the loss of information of the equipment running state represented by the state features, which will make that the extreme value of the MDs based on its calculation is larger and the dispersion of the data increases.



(a) Boxplot of the original space (b) Boxplot of 3σ space (c) Boxplot of 2σ space

Fig. 2 The box plots of three different sets of the MD

4.3 Validity Analysis and Comparison

After determining the threshold based on the normal training samples, the MDs of the other samples were calculated based on the different sample spaces and the indicators of the model robustness were calculated according to Eqs. (8)—(11) (shown in Table 2).

From the analysis of the table, it can be seen that under the premise of using the box plot method to determine the thresholds of the models, the accuracy of the models constructed by the control chart is greatly improved compared with the model of the original space. After screening the extreme data out through the control chart, the models of MS_1 and MS_2 can basically fully recognized the abnormal state of the equipment, and the three models have the same recognition ability for the normal state of the equipment, which makes the RS of the models of MS_1 and MS_2 much higher than the model of the original space. However, it is reasonable to make the models biased towards the high sensitivity to recognize the abnormal state in order to meet the high accuracy requirements to monitor the abnormal running state of the equipment.

Then, in order to further validate the validity of the models based on the control chart and discuss the relationship between the construction of the MS and the determination of the threshold, we measured the original samples based on the MS_1 and MS_2, which is also the practical application requirement on the original data. And in order to compare the threshold determination based on the box plot method with the traditional method[16-22], we used different thresholds to perform the discriminant analysis on the original sample space and further calculated the robustness of the models (shown in Table 3).

Table 3 shows in the use of box plot to build MS conditions, the model threshold is determined by the traditional method which has low accuracy and sensitivity on the original sample space. And at this time, the model is seriously biased towards recognizing the normal running state of the equipment away from recognizing the abnormal running state of the equipment, and that the model is seriously imbalanced, which does not meet the high sensitivity requirements to monitoring the abnormal state. In contrast, the model based on the box plot method to calculate the threshold still

Table 2 The robustness of the three models

Data	Training sample				Test sample			
	Accuracy/%	Sensitivity/%	Specificity/%	RS	Accuracy/%	Sensitivity/%	Specificity/%	RS
MS_0 MD=0.969 568 5	81.95	84.65	75.00	1.13	81.29	84.81	72.40	1.17
MS_1 MD=0.904 630 1	93.03	100.00	74.98	1.33	92.12	100.00	70.99	1.41
MS_2 MD=0.965 970 5	92.81	100.00	74.97	1.33	93.60	100.00	76.19	1.31

Table 3 Robustness of the original sample spaces under MS_1 and MS_2

Data		Training sample				Test sample			
		Accuracy/%	Sensitivity/%	Specificity/%	RS	Accuracy/%	Sensitivity/%	Specificity/%	RS
MS_2	The tranditional MD=4.81556	41.89	24.89	85.57	0.29	40.46	23.42	83.44	0.28
	The box plot method MD=1.335508	92.10	98.76	75.00	1.32	91.52	99.10	72.40	1.37
MS_2	The tranditional MD=13.9215	40.84	15.81	84.14	0.19	29.59	9.27	80.84	0.11
	The box plot method MD=1.882078	92.13	98.79	75.00	1.32	90.69	98.20	71.75	1.37

has high accuracy and sensitivity on the original sample space, which has no significant difference on recognizing the samples in the corresponding control limits.

4.4 State Feature Variable Selection

In the previous section, we compared the robustness and validity between the traditional model and the improved one, and found that the improved model has a strong ability to recognize the running state of the equipment. In this section, there will be the orthogonal design and SNR calculation based on MS_1 and MS_2, so as to select useful features variables. Since there are 11 state feature variables, we chose the OA of two levels ($L_{12}(2^{11})$) to be the orthogonal design, and then according to Eqs. (6) and (7), the gain of SNR is calculated for each feature variable. Finally, we selected seven variables which are 1st, 2nd, 4th, 5th, 9th, 10th and 11th based on the MS_1 and selected six variables including 2nd, 4th, 5th, 9th, 10th and 11th based on the MS_2 (shown as Figs. 3 and 4).

According to the second part of the previous discussion on the state feature selection: the greater the SNR gain or the greater the SNR is, the stronger the ability of the corresponding state feature to recognize and represent the running state of the equipment is. Therefore, we should select those state feature variables which have the higher accuracy to characterize the actual state of the equipment to monitor the running state of the equipment, which will make the utilization of the data to be maximized and reduce the use of sensing equipment to save the corresponding costs.

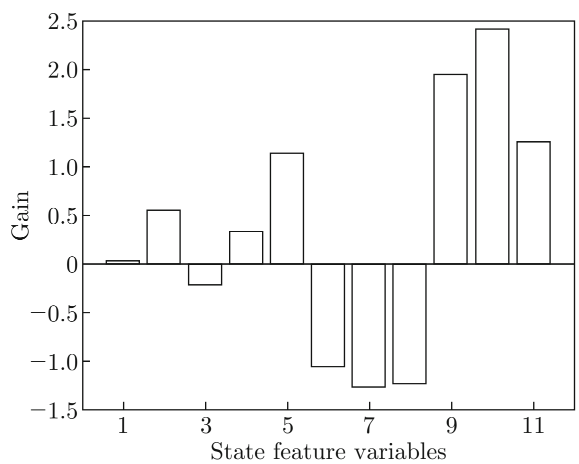


Fig. 3 The SNR gain based on MS_1

At the same time, under the premise of guaranteeing the judgement accuracy, in the actual state feature selection, we should choose as few state feature variables as possible to characterize the running state of the equipment, so as to reduce the calculation time and improve the efficiency of the model to recognize the equipment state. Therefore, based on the size of the gains, we validate the validity of the model based on the abnormal samples through the relative sensitivity and the accuracy (shown in Tables 4 and 5).

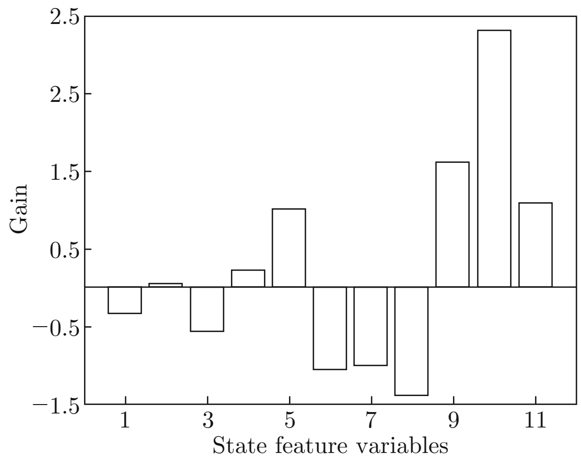


Fig. 4 The SNR gain based on MS_2

Table 4 Selection model evaluation based on MS_1

SRN	Features selection	Threshold	Accuracy/%	RS
Original		1.335 5	92.10	1.32
> 0	1, 2, 4, 5, 9, 10, 11	1.521 0	92.05	1.32
> 0.1	2, 4, 5, 9, 10, 11	1.553 3	88.08	1.24
> 0.4	2, 5, 9, 10, 11	1.540 1	89.37	1.27
> 0.6	5, 9, 10, 11	1.422 4	89.68	1.27
> 1.2	9, 10, 11	1.427 5	90.27	1.28
> 1.5	9, 10	0.936 4	91.11	1.30
> 2.0	10	0.516 7	92.05	1.32

Table 5 Selection model evaluation based on MS_2

SNR	Features selection	Threshold	Accuracy/%	RS
Original		1.882 1	92.13	1.32
> 0	2, 4, 5, 9, 10, 11	2.182 9	89.63	1.27
> 0.1	4, 5, 9, 10, 11	2.201 3	89.07	1.26
> 0.3	5, 9, 10, 11	2.330 1	90.24	1.28
> 1.1	9, 10	1.373 1	91.80	1.31
> 1.7	10	1.510 8	91.98	1.31

From the results in the tables, we can see that with the step by step to improve the size of the gain to remove the corresponding state feature variables, the accuracy of the model shows a trend of decrease first and then increase. Although the trend is not significant, it also shows that different state features have different contributions to the ability of model to recognize the running state of the equipment, and the magnitude of its contribution is not only related to the SNR gain, but also to different combinations of state features and the correlation between different feature variables.

After analysis, we finally chose the combination of the 1st, 2nd, 4th, 5th, 9th, 10th, 11th state features to construct the improved measurement table, which respectively represents the failure frequency of the bearing guard, the frequency of the rolling body through the inner ring, the root mean square of the vibration signal and the temperature of the bearing.

5 Conclusion

MTS has been widely used in pattern recognition and data mining because of its high efficiency in data classification and simplicity. However, the practical application data do not exactly conform to the corresponding statistical distribution, that is to say, the data distribution is irregular. And because of the multidimensional nature of the data, there is a case where the individual dimension data are extremely changing while the overall observation indicates the equipment state is normal. These conditions will have a great impact on the accuracy and stability of the MS and the threshold which is determined based on the normal training samples. And then the performance of the model will be impacted greatly, so that the model cannot correctly characterize and recognize the running state of the equipment. In view of the above problems, this paper presents an improved model based on the control chart and the box plot method. Firstly, the data of each dimension of the training samples are screened by the control chart to remove the influence of extreme change, and traditionally the extreme change is considered to be normal in the range of the original MS. Secondly, the box plots are used to statistically describe the calculated results to analyze the overall distribution and discreteness of the results. And then according to the results, the threshold of the model can be determined based on the upper quartile. Finally, an example is used to compare and analyze the different models. And the result shows that compared with the traditional model, the indices of the improved model have been improved by nearly 20% in the field of the accuracy and sensitivity to recognize the running state of the equipment. And at the same time, while considering the relationship between the MS construction and the threshold determination, the accuracy of the model in the original space increases from 40% to 90% compared with that of the traditional model, and the sensitivity also increases from 20% to 98%. Obviously, the improved model is more accurate and robust with the practical requirement.

References

[1] TAGUCHI G, JUGULUM R. The Mahalanobis-Taguchi strategy: A pattern technology system [M]. Hoboken, NJ, USA: John Wiley & Sons, 2002.

[2] WOODALL W H, KOUDELIK R, TSUI K L, et al. A review and analysis of the Mahalanobis-Taguchi system [J]. *Technometrics*, 2003, **45**(1): 1-15.

[3] YAZID A M, RIJAL J K, AWALUDDIN M S, et al. Pattern recognition on remanufacturing automotive component as support decision making using

Mahalanobis-Taguchi system [J]. *Procedia CIRP*, 2015, **26**: 258-263.

[4] ABU M Y, JAMALUDIN K R, RAMLIE F. Pattern recognition using Mahalanobis-Taguchi system on connecting rod through remanufacturing process: A case study [J]. *Advanced Materials Research*, 2013, **845**: 584-589.

[5] MOHAN D, SAYGIN C, SARANGAPANI J. Real-time detection of grip length deviation during pull-type fastening: a Mahalanobis-Taguchi System (MTS)-based approach [J]. *International Journal of Advanced Manufacturing Technology*, 2008, **39**(9/10): 995-1008.

[6] MAKAJIMA H, TAKADA K, YANO H, et al. Predictive evaluation and efficient management of medical examinations using Mahalanobis Taguchi system method [J]. *Japanese Journal of Public Health*, 1999, **46**(5): 351-363(in Japanese).

[7] DAS P, DATTA S. Exploring the effects of chemical composition in hot rolled steel product using Mahalanobis distance scale under Mahalanobis-Taguchi system [J]. *Computational Materials Science*, 2007, **38**(4): 671-677.

[8] CUDNEY E A, PARYANI K, RAGSDELL K M. Applying the Mahalanobis-Taguchi system to vehicle handling [J]. *Concurrent Engineering*, 2006, **14**(4): 343-354.

[9] ZENG J H, ZENG F Z. The measurement scale of Mahalanobis-Taguchi system optimization based on fuzzy robustness discriminant criterion [J]. *Industrial Engineering and Management*, 2008, **13**(3): 52-55(in Chinese).

[10] NAKATSUGAWA M, OHUCHI A. A study on determination of the threshold in MTS algorithm [J]. *Transactions of the Institute of Electronics Information and Communication Engineers A*, 2001, **84**(4): 519-527.

[11] SU C T, HSIAO Y H. An evaluation of the robustness of MTS for imbalanced data [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, **19**(10): 1321-1332.

[12] XU Q, ZHENG C D, HAN Z J. Study of Mahalanobis-Taguchi system in multiple recognition [J]. *Journal of Nanjing University of Science and Technology*, 2002, **26**(1): 92-95 (in Chinese).

[13] IQUEBAL A S, PAL A, CEGLAREK D, et al. Enhancement of Mahalanobis-Taguchi system via rough sets based feature selection [J]. *Expert Systems with Applications*, 2014, **41**(17): 8003-8015.

[14] FOSTER C R, JUGULUM R, FREY D D. Evaluating an adaptive one-factor-at-a-time search procedure within the Mahalanobis-Taguchi system [J]. *International Journal of Industrial and Systems Engineering*, 2009, **4**(6): 600-614.

[15] RAI B K, CHINNAM R B, SINGH N. Prediction of drill-bit breakage from degradation signals using Mahalanobis-Taguchi system analysis [J]. *International Journal of Industrial and Systems Engineering*, 2008, **3**(2): 134-148.

[16] WANG N, SAYGIN C, SUN S D. Impact of Mahalanobis space construction on effectiveness of Mahalanobis-Taguchi system [J]. *International Journal of Industrial and Systems Engineering*, 2013, **13**(2): 233-249.

[17] LIPARAS D, ANGELIS L, FELDT R. Applying the Mahalanobis-Taguchi strategy for software defect diagnosis [J]. *Automated Software Engineering*, 2012, **19**(2): 141-165.

[18] LIPARAS D, LASKARIS N, ANGELIS L. Incorporating resting state dynamics in the analysis of encephalographic responses by means of the Mahalanobis-Taguchi strategy [J]. *Expert Systems with Applications*, 2013, **40**(7): 2621-2630.

[19] YANG T, CHENG Y T. The use of Mahalanobis-Taguchi System to improve flip-chip bumping height inspection efficiency [J]. *Microelectronics Reliability*, 2010, **50**(3): 407-414.

[20] GHASEMI E, AAGHAIE A, CUDNEY E A. Mahalanobis Taguchi system: A review [J]. *International Journal of Quality & Reliability Management*, 2015, **32**(3): 291-307.

[21] SOYLEMEZOGLU A, JAGANNATHAN S, SAYGIN C. Mahalanobis Taguchi system (MTS) as a prognostics tool for rolling element bearing failures [J]. *Journal of Manufacturing Science and Engineering*, 2010, **132**(5): 051014.

[22] JUGULUM R, TAGUCHI G, TAGUCHI S, et al. A review and analysis of the Mahalanobis-Taguchi system: Discussion [J]. *Technometrics*, 2003, **45**(1): 16-21.

[23] GAUDARD M, RAMSEY P J. Introduction to statistical quality control [J]. *Technometrics*, 1997, **39**(3): 331-332.

[24] BENJAMINI Y. Opening the box of a boxplot [J]. *The American Statistician*, 1988, **42**(4): 257-262.

[25] SU C T, HSIAO Y H. An evaluation of the robustness of MTS for imbalanced data [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, **19**(10): 1321-1332.