Hindawi
Computational Intelligence and Neuroscience
Volume 2020, Article ID 4609423, 15 pages
https://doi.org/10.1155/2020/4609423

Research Article

Optimized Mahalanobis–Taguchi System for High-Dimensional Small Sample Data Classification

Xinping Xiao, Dian Fu, Yu Shi, and Jianghui Wen

School of Science, Wuhan University of Technology, Wuhan 430070, China

Correspondence should be addressed to Dian Fu; dianfu@whut.edu.cn

Received 21 September 2019; Revised 15 December 2019; Accepted 28 December 2019; Published 26 April 2020

Academic Editor: José Alfredo Hernández-Pérez

The Mahalanobis–Taguchi system (MTS) is a multivariate data diagnosis and prediction technology, which is widely used to optimize large sample data or unbalanced data, but it is rarely used for high-dimensional small sample data. In this paper, the optimized MTS for the classification of high-dimensional small sample data is discussed from two aspects, namely, the inverse matrix instability of the covariance matrix and the instability of feature selection. Firstly, based on regularization and smoothing techniques, this paper proposes a modified Mahalanobis metric to calculate the Mahalanobis distance, which is aimed at reducing the influence of the inverse matrix instability under small sample conditions. Secondly, the minimum redundancy-maximum relevance (mRMR) algorithm is introduced into the MTS for the instability problem of feature selection. By using the mRMR algorithm and signal-to-noise ratio (SNR), a two-stage feature selection method is proposed: the mRMR algorithm is first used to remove noise and redundant variables; the orthogonal table and SNR are then used to screen the combination of variables that make great contribution to classification. Then, the feasibility and simplicity of the optimized MTS are shown in five datasets from the UCI database. The Mahalanobis distance based on regularization and smoothing techniques (RS-MD) is more robust than the traditional Mahalanobis distance. The two-stage feature selection method improves the effectiveness of feature selection for MTS. Finally, the optimized MTS is applied to email classification of the Spambase dataset. The results show that the optimized MTS outperforms the classical MTS and the other 3 machine learning algorithms.

1. Introduction

The Mahalanobis–Taguchi system (MTS) uses the Mahalanobis distance (MD) as a measurement scale and combines Taguchi robust design to achieve system diagnosis and dimension optimization. The MTS is a commonly used multisystem pattern recognition method, which has achieved good results in medical diagnosis [1, 2], financial early warning [3], product detection [4, 5], fault analysis [6], enterprise management, comprehensive evaluation [7], and so on. The MTS is widely applied to the optimization and classification of large sample data or imbalanced data [6, 8–12]. However, in the field of pattern recognition, a large number of recognition problems belong to high-dimensional small sample size problem, and the research on high-dimensional small sample size problem has gradually become a hot spot. For example, image analysis is a typical high-dimensional small sample problem in the field of pattern recognition, and it is also the focus of machine vision field. Image processing is of great significance for machine vision issue and image analysis. The existence of noise in some images makes image processing difficult. In the mentioned images, a speckle noise is the main problem which severely damages the image because of its multiplicative property. Simultaneously, existence of speckle, clutter edge, and image-level clutters can also make false alarms and false detections on retrieval algorithms. It can effectively improve the recognition ability through denoising. Deep learning has a good application in the field of image denoising algorithms. For example, the supervised deep learning method based on deep belief network (DBN) was used to detect changes in synthetic aperture radar (SAR) images [13], a neural network with hybrid algorithm of CNN and multilayer perceptron (CNN-MLP) was

suggested for image classification [14], and so on. In addition, image segmentation is the key step from image processing to image analysis. Effective image segmentation method can improve the recognition effect of machine vision [15–17]. Image segmentation is also an important step in texture recognition of images, and the existence of speckle noise will affect image segmentation. Therefore, an algorithm based on wavelet transform and support vector machine was proposed for texture recognition of SAR images [18]. Image processing technology has important applications in image registration [19], coastline detection [20, 21], and so on. Therefore, studying the MTS for high-dimensional small sample data not only provides new ideas for dimension reduction and classification of small sample problems but also extends the application range of the MTS so that the MTS can also play a role in intelligent traffic system [22], image processing, machine vision, and other techniques of electronics field.

The present research on high-dimensional small sample data mainly focuses on three aspects. First, the number of training samples is smaller than that of variables, which will cause the singularity problem of the covariance matrix. Second, when the number of training samples is slightly larger than that of variables, biased eigenvalue estimation will cause the inverse matrix instability of the covariance matrix. Third, the feature selection problem occurs. For the case in which the number of training samples is less than that of features, a common method is to increase the sample size by generating a virtual sample. By using the Monte Carlo method, Karaivanova et al. [23] reconstructed the probability distribution of insufficient data to generate virtual samples. Based on virtual sample generation technology, Gong et al. [24] proposed a new particle swarm optimization (PSO) algorithm to generate effective virtual samples. For the case in which the number of training samples is slightly larger than that of variables, the covariance matrix is optimized mainly from the perspective of eigenvalues. With regard to the poor learning performance of the keep it simple and straightforward (KISS) metric, Tao et al. [25] proposed a double regularization KISS metric learning method for pedestrian recognition problem. Through adjusting the eigenvalues of intraclass and interclass covariance matrices according to the discriminant information of training samples, Liong et al. [26] proposed a new discriminant regularization metric learning method to minimize the estimated distance metric deviation. For the feature selection problem, the classification performance and stability are discussed. Espezua et al. [27] compressed data rapidly and then used an improved projection tracking method to avoid dimension disasters. Hira and Gillies [28] summarized various methods about dimension reduction for high-dimensional microarray data. Kamyab and Eftekhari [29] used a multimodal optimization technology to solve feature selection problems. Goh and Wong [30] proposed a sort-based network algorithm for feature selection in proteomics. To improve the effectiveness and robustness of feature selection technology, Du et al. [31] proposed a hybrid feature selection method based on multicore learning. These methods indicate that the feature selection for high-dimensional data should not only consider the classification performance but also ensure the stability of the results. These studies mainly focused on covariance matrix and feature selection, and few methods can simultaneously solve the problem of dimension reduction and classification for high-dimensional small sample data. Unlike most classification methods, the MTS can screen effective features and construct a classification model by determining threshold. Hence, the MTS can simultaneously solve the problem of dimension reduction and classification.

Many studies have also focused on covariance matrix and feature selection for the MTS. For the covariance matrix, when multiple collinearities occur among variables, the inverse matrix of the covariance matrix does not exist. Taguchi used the Schmidt orthogonalization [32] and the adjoint matrix [33] to calculate MD. Based on Schmidt orthogonalization, Su and Hsiao [34] proposed weighted Schmidt orthogonalization to calculate MD. Shakya et al. [35] used an integrated Schmidt orthogonalization method for the classification of rolling bearings. On the basis of the generalized inverse matrix, Han et al. [36] redefined MD and proposed the Mahalanobis–Taguchi generalized inverse matrix method. Chang et al. [37] used the pseudoinverse of the covariance matrix to calculate MD. Through eliminating multicollinearity by the ridge estimation method, Tao and Cheng [38] proposed the ridge-MD that combines the ridge estimate with MD. For the feature selection, the classical MTS uses the orthogonal table and signal-to-noise ratio (SNR) methods to screen variables. Abraham and Variyath [39] confirmed the possibility of using appropriate algorithms for the dimension reduction and optimization of the MTS. Reséndiz et al. [40] applied the binary ant colony optimization algorithm to optimize the variable combination. Iquebal et al. [41] screened variables on the basis of the maximized degree of dependence between variables and classes or among categories. Reséndiz-Flores et al. [42] used the hybrid binary heuristic algorithm of PSO and gravity search algorithm for feature selection. By introducing chaos mapping and binary particle swarm optimization algorithm, Gu et al. [3] constructed an improved MTS-CBPSO method to screen effective variables. Reyes-Carlos et al. [43] constructed the mathematical model to select features and used metaheuristic algorithms to solve the corresponding model. To solve the feature selection problem of unbalanced welding data, Mahmoud et al. [44] applied the genetic algorithm to the MTS and proposed the Mahalanobis genetic algorithm classifier. Niu and Cheng [45] used optimization model to select variables and constructed probability threshold model for unbalanced data classification. Most of these studies focused on large sample data or unbalanced data, whereas few studies discussed the high-dimensional small sample data. For the covariance matrix, existing research only discussed the multiple collinearity among variables, but few studies discussed the inverse matrix instability of the covariance matrix under the condition of small sample data. For the feature selection, existing research mainly screened

features in terms of classification performance, but few studies discussed the instability of feature selection.

In the current work, the optimized MTS for the classification of high-dimensional small sample data is discussed from two aspects, namely, the inverse matrix instability of the covariance matrix and the instability of feature selection. Aimed at the inverse matrix instability problem of the covariance matrix, the Mahalanobis metric based on regularization [25, 46] and smoothing [47, 48] techniques is proposed. Aimed at the instability problem of feature selection, a two-stage feature selection method based on the minimum redundancy-maximum relevance (mRMR) [49, 50] feature selection algorithm and SNR is proposed.

The remainder of this paper is structured as follows. In Section 2, we briefly introduce the implementation steps of the MTS. In Section 3, we construct an optimized MTS model. In Section 4, we select datasets for verification and analysis. In Section 5, we conduct empirical research on the email filtering problem. In Section 6, we derive the conclusion.

2. Mahalanobis–Taguchi System

The MTS is a pattern recognition technology based on MD and the Taguchi experiment design. The initial research of the MTS is a two-classification problem. One is set as the normal observations and the other is set as the abnormal observations. To achieve the purpose of system diagnosis and dimension reduction optimization, the orthogonal table and SNR are used to screen the variables, and the classification threshold is determined in accordance with MD.

Assuming the number of the normal observations is n and the number of the abnormal observations is m, both normal and abnormal observations consist of p variables. The i^{th} observation of the normal observations after normalization is recorded as $Z_i = [z_{i1}, z_{i2}, \ldots, z_{ip}]_T$, $i = 1, 2, \ldots, n$. The abnormal observations are normalized in accordance with the mean and variance of the normal observations, and the i^{th} observation of the abnormal observations is recorded as $Z_i = [z_{i1}, z_{i2}, \ldots, z_{ip}]_T$, $i = n+1, n+2, \ldots, n+m$. The MD from each observation to the reference space can be expressed as

$$\begin{aligned} \mathrm{MD}_i &= \frac{1}{p} Z_i^T \widehat{\Sigma}_0^{-1} Z_i, \quad i = 1, 2, \ldots, n, n+1, \ldots, n+m, \\ \widehat{\Sigma}_0 &= \frac{1}{n-1} \sum_{i=1}^{n} Z_i Z_i^T, \quad i = 1, 2, \ldots, n, \end{aligned} \tag{1}$$

where $\widehat{\Sigma}_0$ is the covariance matrix of the normal observations. In the Mahalanobis–Taguchi system, the standardized variables of the normal observations are used to construct a reference space. The MD of the abnormal observations is significantly larger than that of the normal observations, indicating that the constructed reference space is valid; otherwise, the normal observations should be recollected until a valid reference space is obtained.

The calculation of the traditional MD requires the number of observations is larger than that of variables. Simultaneously, multiple collinearity should be absent among variables to avoid the situation where the inverse matrix of the covariance matrix does not exist. In addition, MD exaggerates the role of variables with minor changes and is susceptible to the instability of the covariance matrix. Therefore, the singularity and instability of the covariance matrix affect the calculation of the traditional MD.

The corresponding two-level orthogonal table $L_r(2^p)$, where r represents the number of trials, is selected in accordance with the number of initial p variables. The level of "1" indicates that the variable is selected, and the level of "2" indicates that the variable is not selected. On the basis of the information of the orthogonal table, the reference space is reconstructed by using the selected variables for each experiment. The MD of each abnormal observation in the new reference space is calculated, and the larger-the-better SNR is calculated as the response value. On the basis of the idea of the experimental design, the effects at different levels of each variable are analyzed, and effective variables are selected. According to the selected variable combination, the MD of each observation is recalculated to determine the threshold by minimizing the classification loss. The unknown observations are then diagnosed.

The classical MTS uses the orthogonal table and SNR to screen variables and select variable combination with high SNR. To get better classification results, the number of training samples should be sufficient to fully reflect the information of each variable. Otherwise, the selected variable combination will not exert a good classification effect on test samples.

3. Optimized MTS

This section constructs an optimized MTS for high-dimensional small sample data classification. Firstly, based on the regularization and smoothing techniques, the calculation of the modified Mahalanobis metric is introduced, and the feasibility of the modified Mahalanobis metric is proved. Then, based on the mRMR algorithm and SNR, the implementation steps of the two-stage feature selection method are introduced. Finally, the algorithm flow of the optimized Mahalanobis–Taguchi system is introduced.

3.1. Mahalanobis Metric Based on Regularization and Smoothing Techniques. When the number of samples is slightly larger than that of variables, biased eigenvalue estimation will cause the inverse matrix instability of the covariance matrix. Tao et al. [25] proved that the estimation of the covariance matrix is affected by sample size. A small sample size leads to a large generalization bound of covariance matrix estimation. Specifically, the large eigenvalues of the real covariance matrix are overestimated, whereas the small eigenvalues are underestimated. Overestimated large eigenvalues and underestimated small eigenvalues are detrimental to subsequent classification. The calculation of MD depends on the covariance matrix. If the estimation of the covariance matrix is affected, the calculation of MD will produce a deviation. Therefore, the

traditional MD is no longer applicable to the high-dimensional small sample data. Because of the one-to-one correspondence between the covariance matrix and a set of eigenvalues or eigenvectors, the estimation performance of the covariance matrix can be improved by improving the estimation of eigenvalues and eigenvectors. On the basis of the above analysis, regularization and smoothing techniques are introduced to improve the performance of covariance estimation in Mahalanobis metric learning.

The correlation coefficient matrix among p variables of normal observations is a semipositive matrix, which can be expressed as

$$\widehat{\Sigma}_0 = \Phi_0 \Lambda_0 \Phi_0^T, \tag{2}$$

where $\Lambda_0 = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, with λ_i, $i = 1, 2, \ldots, p$ being the i^{th} eigenvalue of $\widehat{\Sigma}_0$, $\Phi_0 = [\phi_1, \phi_2, \ldots, \phi_p]$, with ϕ_i being the eigenvector corresponding to λ_i, and Φ_0 is an orthogonal matrix.

3.1.1. Smoothing Technique. The basic idea of data smoothing technology is to increase low probability, reduce high probability, and make the probability distribution tend to average. The smoothing technique is introduced to eliminate zero eigenvalues and make the distribution of eigenvalues smooth. However, when all the eigenvalues tend to be the same, the information of the original sample is lost. Therefore, the smoothing technique is used to adjust the small eigenvalues of the covariance matrix.

In accordance with the smoothing technique, a small constant β_0 is used to replace the small eigenvalues of the covariance matrix, which is recorded as

$$\Lambda_1 = \operatorname{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_k, \underbrace{\beta_0 \ldots \beta_0}_{p-k} \right), \quad k = 0, 1, \ldots, p-1, \tag{3}$$

$$\beta_0 = \frac{1}{p-k} \sum_{j=k+1}^{p} \lambda_j. \tag{4}$$

When the smoothing technology is introduced, some small eigenvalues are replaced with average value. This method not only avoids the appearance of zero eigenvalues but also smooths the distribution of eigenvalues.

3.1.2. Regularization Technique. The basic idea of regularization technology is to use a unit matrix to interpolate the covariance matrix; hence, the sample covariance matrix tends to the unit matrix, which is expressed as

$$\begin{aligned} \widehat{\Sigma}_\gamma &= (1-\gamma)\widehat{\Sigma}_0 + \gamma\alpha I = (1-\gamma)\Phi_0 \Lambda_0 \Phi_0^T + \gamma\alpha \Phi_0 \Phi_0^T \\ &= \Phi_0 \left[(1-\gamma)\Lambda_0 + \gamma\alpha I \right] \Phi_0^T, \end{aligned} \tag{5}$$

where $\alpha = (1/p)\operatorname{tr}(\widehat{\Sigma}_0)$, $0 < \gamma < 1$.

After the introduction of the regularization technique, the eigenvalue corresponding to the covariance matrix becomes

$$\begin{aligned} \Lambda_2 = (1-\gamma)\Lambda_0 + \gamma\alpha I = \operatorname{diag}\big(&(1-\gamma)\lambda_1 + \gamma\alpha, (1-\gamma)\lambda_2 \\ &+ \gamma\alpha, \ldots, (1-\gamma)\lambda_p + \gamma\alpha \big). \end{aligned} \tag{6}$$

The large eigenvalues of the original covariance matrix decrease because of the existence of parameter γ. Therefore, parameter γ can make $\widehat{\Sigma}_0$ tend to the unit matrix and restrain the overestimation of large eigenvalues.

3.1.3. Mahalanobis Metric Based on Regularization and Smoothing Techniques. For a limited training samples, the estimation of the covariance matrix produces deviations, and the calculation of the traditional MD is affected. In view of the one-to-one correspondence between the covariance matrix and eigenvalues or eigenvectors, the performance of the covariance matrix can be improved by adjusting the eigenvalues, that is, reducing the overestimated large eigenvalues and increasing the underestimated small eigenvalues. Regularization and smoothing technologies are thus introduced into the calculation of MD under the condition of limited samples. Smoothing technology is used to improve the estimation of small eigenvalues, and regularization technology is used to reduce the influence of overestimated large eigenvalues.

The sample covariance matrix is processed by regularization and smoothing techniques, and the new estimation is as follows:

$$\widehat{\Sigma}'_{\gamma,k} = \Phi_0 \Lambda_{\gamma,k} \Phi_0^T, \tag{7}$$

where

$$\begin{aligned} \Lambda_{\gamma,k} = \operatorname{diag}\big(&(1-\gamma)\lambda_1 + \gamma\alpha, (1-\gamma)\lambda_2 + \gamma\alpha, \ldots, (1-\gamma)\lambda_k \\ &+ \gamma\alpha, (1-\gamma)\beta_0 + \gamma\alpha, \ldots, (1-\gamma)\beta_0 + \gamma\alpha \big). \end{aligned} \tag{8}$$

The calculation of Mahalanobis distance based on regularization and smoothing techniques (RS-MD) of each sample is transformed into

$$\mathrm{MD}'_i = \frac{1}{p} Z_i^T \widehat{\Sigma}'^{-1}_{\gamma,k} Z_i, \quad i = 1, 2, \ldots, n. \tag{9}$$

Theorem 1. The observation is assumed to have an upper bound, that is, $\forall Z_i \in Z$, $\|Z_i\| \le \Omega_Z$. For any two samples Z_i and Z_j standardized in the same category, we have

$$\left| \mathrm{MD}'_i - \mathrm{MD}'_j \right| \le \frac{2}{p} \Omega_Z^2 \sqrt{ \sum_{i=1}^{p} \frac{1}{\left(\left(\Lambda_{\gamma,k} \right)_{ii} \right)^2} }, \tag{10}$$

where $(\Lambda_{\gamma,k})_{ii}$ represents the i^{th} diagonal element of $\Lambda_{\gamma,k}$.

Proof. From equation (9),

$$\begin{aligned}\left|\mathrm{MD}_i' - \mathrm{MD}_j'\right| &= \left|\frac{1}{p}Z_i^{\mathrm{T}}\widehat{\Sigma}_{\gamma,k}'^{-1}Z_i - \frac{1}{p}Z_j^{\mathrm{T}}\widehat{\Sigma}_{\gamma,k}'^{-1}Z_j\right| = \frac{1}{p}\left|\mathrm{trace}\left(Z_i^{\mathrm{T}}\widehat{\Sigma}_{\gamma,k}'^{-1}Z_i\right) - \mathrm{trace}\left(Z_j^{\mathrm{T}}\widehat{\Sigma}_{\gamma,k}'^{-1}Z_j\right)\right| \\ &= \frac{1}{p}\left|\mathrm{trace}\left(\widehat{\Sigma}_{\gamma,k}'^{-1}Z_iZ_i^{\mathrm{T}}\right) - \mathrm{trace}\left(\widehat{\Sigma}_{\gamma,k}'^{-1}Z_jZ_j^{\mathrm{T}}\right)\right| = \frac{1}{p}\left|\mathrm{trace}\left[\widehat{\Sigma}_{\gamma,k}'^{-1}\left(Z_iZ_i^{\mathrm{T}} - Z_jZ_j^{\mathrm{T}}\right)\right]\right| \\ &\le \frac{1}{p}\left|\mathrm{trace}\left(\left\|\widehat{\Sigma}_{\gamma,k}'^{-1}\right\|\left\|Z_iZ_i^{\mathrm{T}} - Z_jZ_j^{\mathrm{T}}\right\|\right)\right| = \frac{1}{p}\left\|\widehat{\Sigma}_{\gamma,k}'^{-1}\right\|\left\|Z_iZ_i^{\mathrm{T}} - Z_jZ_j^{\mathrm{T}}\right\|.\end{aligned} \tag{11}$$

Given $\|Z_i\| \le \Omega_Z$, we determine

$$\begin{aligned}\left\|Z_iZ_i^{\mathrm{T}}\right\| &\le \left\|Z_i\right\|\left\|Z_i^{\mathrm{T}}\right\| = \Omega_Z^2, \\ \left\|Z_iZ_i^{\mathrm{T}} - Z_jZ_j^{\mathrm{T}}\right\| &\le \left\|Z_i\right\|\left\|Z_i^{\mathrm{T}}\right\| + \left\|Z_j\right\|\left\|Z_j^{\mathrm{T}}\right\| = 2\Omega_Z^2.\end{aligned} \tag{12}$$

With

$$\begin{aligned}\left\|\widehat{\Sigma}_{\gamma,k}'^{-1}\right\| &= \sqrt{\mathrm{trace}\left[\left(\widehat{\Sigma}_{\gamma,k}'^{-1}\right)\left(\widehat{\Sigma}_{\gamma,k}'^{-1}\right)^T\right]} = \sqrt{\mathrm{trace}\left[\left(\Phi_0\Lambda_{\gamma,k}^{-1}\Phi_0^T\right)\left(\Phi_0\Lambda_{\gamma,k}^{-1}\Phi_0^T\right)^T\right]} = \sqrt{\mathrm{trace}\left[\Phi_0\left(\Lambda_{\gamma,k}^{-1}\right)^2\Phi_0^T\right]} \\ &= \sqrt{\mathrm{trace}\left[\left(\Lambda_{\gamma,k}^{-1}\right)^2\Phi_0^T\Phi_0\right]} = \sqrt{\mathrm{trace}\left(\Lambda_{\gamma,k}^{-1}\right)^2} = \sqrt{\sum_{i=1}^{p}\frac{1}{\left(\left(\Lambda_{\gamma,k}\right)_{ii}\right)^2}},\end{aligned} \tag{13}$$

we yield

$$\left|\mathrm{MD}_i' - \mathrm{MD}_j'\right| \le \frac{2}{p}\Omega_Z^2\sqrt{\sum_{i=1}^{p}\frac{1}{\left(\left(\Lambda_{\gamma,k}\right)_{ii}\right)^2}}. \tag{14}$$

Theorem 1 indicates that for any two samples from the same class, the upper bound of the difference of RS-MD or MD is related to the eigenvalue of the covariance matrix. Adjusting eigenvalues can improve the performance of the covariance matrix and thus improve the robustness of MD.

Theorem 2. Let $(\Lambda_0)_{ii}$, $(\Lambda_1)_{ii}$, and $(\Lambda_2)_{ii}$ denote the i^{th} diagonal elements of Λ_0, Λ_1, and Λ_2, respectively,

$$\begin{aligned}\Psi_0 &= \sum_{i=1}^{p}\frac{1}{\left(\left(\Lambda_0\right)_{ii}\right)^2}, \\ \Psi_1 &= \sum_{i=1}^{p}\frac{1}{\left(\left(\Lambda_1\right)_{ii}\right)^2}, \\ \Psi_2 &= \sum_{i=1}^{p}\frac{1}{\left(\left(\Lambda_2\right)_{ii}\right)^2}, \\ \Psi_{\gamma,k} &= \sum_{i=1}^{p}\frac{1}{\left(\left(\Lambda_{\gamma,k}\right)_{ii}\right)^2},\end{aligned} \tag{15}$$

and then

$$\begin{aligned}\Psi_1 &\le \Psi_0, \\ \Psi_{\gamma,k} &\le \Psi_2 \le \Psi_0.\end{aligned} \tag{16}$$

Proof. The eigenvalue of the sample covariance matrix becomes equation (3) by processing with the smoothing technique. Then,

$$\begin{aligned}\Psi_1 &= \sum_{i=1}^{p}\frac{1}{\left(\left(\Lambda_1\right)_{ii}\right)^2} = \sum_{i=1}^{k}\frac{1}{\lambda_i^2} + \sum_{i=k+1}^{p}\frac{1}{\beta_0^2} = \sum_{i=1}^{k}\frac{1}{\lambda_i^2} \\ &\quad + \sum_{i=k+1}^{p}\left(\frac{1}{(1/p-k)\sum_{i=k+1}^{p}\lambda_i}\right)^2 \\ &\le \sum_{i=1}^{k}\frac{1}{\lambda_i^2} + \sum_{i=k+1}^{p}\left(\frac{(p-k)^2}{(1/p-k)\sum_{i=k+1}^{p}\lambda_i}\right)^2 \\ &= \sum_{i=1}^{k}\frac{1}{\lambda_i^2} + \sum_{i=k+1}^{p}\left(\frac{(p-k)^2/\sum_{i=k+1}^{p}\lambda_i}{(p-k)}\right)^2 \\ &\le \sum_{i=1}^{k}\frac{1}{\lambda_i^2} + \sum_{i=k+1}^{p}\left(\frac{\sum_{i=k+1}^{p}(1/\lambda_i)}{(p-k)}\right)^2 \\ &= \sum_{i=1}^{k}\frac{1}{\lambda_i^2} + (p-k)\left(\frac{\sum_{i=k+1}^{p}(1/\lambda_i)}{(p-k)}\right)^2 \\ &\le \sum_{i=1}^{k}\frac{1}{\lambda_i^2} + \sum_{i=k+1}^{p}\frac{1}{\lambda_i^2} = \sum_{i=1}^{p}\frac{1}{\lambda_i^2} = \sum_{i=1}^{p}\frac{1}{\left(\left(\Lambda_0\right)_{ii}\right)^2} = \Psi_0.\end{aligned} \tag{17}$$

The eigenvalue of the sample covariance matrix becomes equation (6) by processing with the regularization technique. Then,

$$\Psi_2 = \sum_{i=1}^{p}\frac{1}{\left(\left(\Lambda_2\right)_{ii}\right)^2} = \sum_{i=1}^{p}\frac{1}{\left[(1-\gamma)\lambda_i + \gamma\alpha\right]^2}. \tag{18}$$

We assume that

$$g(\gamma) = \sum_{i=1}^{p} \frac{1}{\left[(1-\gamma)\lambda_i + \gamma\alpha\right]^2}, \quad 0 < \gamma < 1,$$
$$\alpha = \left(\frac{1}{p}\right)\mathrm{tr}(\widehat{\Sigma}), \tag{19}$$

which yields

$$g'(\gamma) = \sum_{i=1}^{p} \frac{-2(\alpha - \lambda_i)}{\left[\lambda_i + (\alpha - \lambda_i)\gamma\right]^3},$$
$$g''(\gamma) = \sum_{i=1}^{p} \frac{6(\alpha - \lambda_i)^2}{\left[\lambda_i + (\alpha - \lambda_i)\gamma\right]^4} > 0. \tag{20}$$

Accordingly, $g'(\gamma)$ is monotonically increasing in $\gamma \in (0,1)$, and

$$g'(\gamma) < g'(1) = \sum_{i=1}^{p} \frac{-2(\alpha - \lambda_i)}{\left[\lambda_i + (\alpha - \lambda_i)\gamma\right]^3}\bigg|_{\gamma=1} = \sum_{i=1}^{p} \frac{-2(\alpha - \lambda_i)}{\alpha^3} = 0. \tag{21}$$

By contrast, $g(\gamma)$ is monotonically decreasing in $\gamma \in (0,1)$, and

$$\Psi_2 = g(\gamma) < g(0) = \sum_{i=1}^{p} \frac{1}{\lambda_i^2} = \sum_{i=1}^{p} \frac{1}{\left((\Lambda_0)_{ii}\right)^2} = \Psi_0. \tag{22}$$

The eigenvalue of the sample covariance matrix becomes equation (8) by processing with the regularization and smoothing techniques. Then,

$$\Psi_{\gamma,k} = \sum_{i=1}^{p} \frac{1}{\left(\left(\Lambda_{\gamma,k}\right)_{ii}\right)^2} = \sum_{i=1}^{k} \frac{1}{\left[(1-\gamma)\lambda_i + \gamma\alpha\right]^2} + \sum_{i=k+1}^{p} \frac{1}{\left[(1-\gamma)\beta_0 + \gamma\alpha\right]^2}$$
$$\leq \sum_{i=1}^{k} \frac{1}{\left[(1-\gamma)\lambda_i + \gamma\alpha\right]^2} + \sum_{i=k+1}^{p} \frac{1}{\left[(1-\gamma)\lambda_i + \gamma\alpha\right]^2}$$
$$= \sum_{i=1}^{p} \frac{1}{\left[(1-\gamma)\lambda_i + \gamma\alpha\right]^2} = \Psi_2. \tag{23}$$

Thus,

$$\Psi_1 \leq \Psi_0,$$
$$\Psi_{\gamma,k} \leq \Psi_2 \leq \Psi_0. \tag{24}$$

Theorem 2 reflects the relationship between the calculated value of each eigenvalue sequence when the eigenvalues are processed by different methods. Combining Theorems 1 and 2, we know that for any two samples in the same class, the upper bound of the difference fluctuation of RS-MD is smaller than that of traditional MD. Hence, the Mahalanobis metric based on regularization and smoothing techniques is more robust than the traditional Mahalanobis metric. □

3.2. Two-Stage Feature Selection Algorithm Based on mRMR Algorithm and SNR. High-dimensional small sample data can cause the instability problem of feature selection. When the training samples produce a small disturbance, the selected variable combination may produce a large difference. High-dimensional small sample data often contain a large number of redundant, uncorrelated, and noise features. They cannot fully reflect the feature information due to the small number of training samples, thereby resulting in great differences in the selection of feature combination for different training samples. The MTS screens variables only from the perspective of classification accuracy. However, for limited training samples, the selected variable combination based on classification accuracy is no longer reliable. This paper accordingly proposes a two-stage feature selection method based on the mRMR algorithm and SNR. First, the mRMR algorithm is used to remove redundant and noise features, and the feature which is highly relevant to class labels is selected. Then, in accordance with the orthogonal table and SNR, a feature subset with strong resolution is selected to achieve the goals of robust optimization and dimension reduction.

3.2.1. One-Time Feature Selection Based on mRMR Algorithm. High-dimensional small sample data contain a large number of redundant, uncorrelated, and noisy features, which not only increase computational complexity significantly and reduce the performance of the classifier but also cause the instability of feature selection. Therefore, the mRMR algorithm is introduced to ensure the validity of the selected features.

According to the cost functions of information difference and information entropy, the mRMR algorithm is aimed at measuring the maximal sample information and minimal relevance among features. The correlation between features and categories or features is measured by mutual information [51]. Mutual information is a measure of the degree of interdependence between two random variables. Extensive mutual information between two random variables indicates a strong correlation between them [52].

The number of samples n, the number of features p, and category c of dataset X are given. The features are recorded as $a_1, a_2, \ldots, a_p$. The value range of feature a_i is V_i, and the value range of category c is V_c.

The mutual information $I(a_i, c)$ between feature a_i and category c is

$$I(a_i, c) = \sum_{v_i \in V_i} \sum_{v_c \in V_c} p(v_i, v_c) \log \frac{p(v_i, v_c)}{p(v_i)p(v_c)}, \tag{25}$$

where $p(v_i, v_c)$ represents the probability that the value of feature a_i is v_i and the value of class c is v_c. A large value of $I(a_i, c)$ shows a high degree of association between feature a_i and category c [53].

The mutual information $I(a_i, a_j)$ between features a_i and a_j is

$$I(a_i, a_j) = \sum_{v_i \in V_i} \sum_{v_j \in V_j} p(v_i, v_j) \log \frac{p(v_i, v_j)}{p(v_i) p(v_j)}, \tag{26}$$

where $p(v_i, v_j)$ represents the probability that the value of feature a_i is v_i and the value of feature a_j is v_j. A large value of $I(a_i, a_j)$ implies a high similarity of feature a_i to feature a_j [53].

The maximum correlation and minimum redundancy of the mRMR algorithm are calculated as follows:

$$\begin{aligned} \max D(V, c):\ D &= \frac{1}{|V|} \sum_{a_i \in V} I(a_i, c), \\ \min R(V):\ R &= \frac{1}{|V|^2} \sum_{a_i, a_0 \in V} I(a_i, a_0). \end{aligned} \tag{27}$$

where V and $|V|$ represent the feature subset and its dimension, respectively; D represents the mean of the mutual information; and R represents the mutual information among features [49].

The mRMR algorithm generates features with minimum redundancy and maximum correlation through the following two criteria:

$$\begin{aligned} &\max \Phi_1(D, R):\ \Phi_1 = D - R, \\ &\max \Phi_2(D, R):\ \Phi_2 = D/R. \end{aligned} \tag{28}$$

3.2.2. Secondary Feature Selection Based on SNR. The mRMR algorithm removes the redundant and noisy features, and the robustness of feature selection is guaranteed. However, the use of this algorithm does not mean that the features in the obtained feature subset are beneficial to the classification. The features that make great contribution to the classification accuracy is further filtered by using orthogonal table and SNR.

A suitable two-level orthogonal table is selected on the basis of the selected feature subset by the mRMR algorithm. According to the information of the orthogonal table, the reference space is reconstructed by using the selected features for each experiment, and the RS-MD of each abnormal observation is calculated in accordance with equation (9). At this point, the calculation of the larger-the-better SNR is as follows:

$$\mathrm{SN} = -10 \lg \left(\frac{1}{m} \sum_{i=n+1}^{n+m} \frac{1}{\mathrm{MD}_i'} \right). \tag{29}$$

For variable x_j, $\overline{\mathrm{SN}_j^+}$ is used to represent the SNR mean when this variable is used; $\overline{\mathrm{SN}_j^-}$ is used to represent the SNR mean when this variable is not used; and $\Delta = \overline{\mathrm{SN}_j^+} - \overline{\mathrm{SN}_j^-}$ represents the SNR increment. When the increment is positive, variable x_j is retained; otherwise, variable x_j is removed. The contribution degree of each variable to the classification accuracy is judged on the basis of the increment in SNR, and the feature combination with a large contribution degree is selected.

The two-stage feature selection not only ensures the robustness of the selected feature combination by using the mRMR algorithm but also improves the classification accuracy by using the orthogonal table and SNR. Therefore, it achieves the goals of robust optimization and dimension reduction The optimized Mahalanobis–Taguchi system uses the Mahalanobis distance based on regularization and smoothing techniques (RS-MD) as a measurement scale and uses the two-stage feature selection method to screen features. The algorithm flow of the optimized Mahalanobis–Taguchi system is presented in Algorithm 1.

4. Effectiveness Verification of Optimized MTS

To verify the robustness of the RS-MD and the validity of the two-stage feature selection method, five datasets from the UCI database are shown in this section. The MTS uses normal observations to construct the reference space, and the information of the reference space is used to calculate the covariance matrix and MD. To satisfy the characteristics of high-dimensional small sample data, the number of samples cannot exceed 10 times the number of features. Data processing is conducted on the selected five datasets to remove missing values and undifferentiated variables. The information obtained is shown in Table 1.

4.1. Comparative Analysis of Traditional MD and RS-MD. The calculation of traditional Mahalanobis distance requires that the covariance matrix is not singular, that is, the number of normal observations is larger than that of features. According to the information of normal observations in Ionosphere, Z-Alizadeh Sani, Parkinson dataset with replicated acoustic features, and Breast Cancer Wisconsin (prognostic) datasets, the benchmark space is constructed and the data are standardized. The MD of each sample in the above datasets is calculated. Because the RS-MD is affected by parameters β_0 and γ, we choose to smooth the eigenvalues less than 0.01. Different parameters γ are also selected for discussion. When parameter γ is taken as 0.2, 0.5, and 0.9, the RS-MD under each dataset is calculated. The calculation results are shown in Figure 1.

Figure 1 shows the distribution of RS-MD for normal observations of each dataset when parameter γ is taken as 0.2, 0.5, and 0.9. When the parameter γ is 0.2 or 0.5, the fluctuations of the calculated RS-MD are minimal, thereby indicating that the results are highly robust when the parameter γ is small. When the parameter γ is 0.9, the fluctuations of the calculated RS-MD become large, indicating that the robustness of the results is weakened when the parameter γ is large. To further reflect the influence of the parameter γ on the calculation results, the variance of RS-MD when the parameter γ takes different values is shown in Table 2.

Table 2 shows the variance of the RS-MD in the normal observations when parameter γ is taken as 0.2, 0.3, 0.5, and 0.9. When parameter γ is less than 0.5, the variance of the calculated RS-MD is small, indicating that the fluctuation is small. When parameter γ is 0.9, the variance of the calculated RS-MD increases gradually. Hence, when parameter γ approaches 1, the fluctuation of the RS-MD increases and the robustness decreases. This is because the eigenvalues of the estimated covariance matrix at this time are almost equal,

Input: Training dataset X, feature set is S;
Output: Feature subset S', threshold T;
(1) Normalize the data and then calculate the RS-MD of each observation to the reference space;
(2) If the RS-MD of the abnormal observation is significantly larger than that of the normal observation, proceed to the next step; otherwise, recollect the data;
(3) Use the mRMR algorithm to remove redundant and noisy features and select the optimal feature subset S_1;
(4) Construct a two-level orthogonal table in accordance with the feature subset S_1. Calculate the SNR in accordance with the RS-MD of the abnormal observation and combine the orthogonal table and the SNR to select the feature subset S' with large contribution;
(5) Recalculate the MD of each sample according to the feature subset S' and then determine the classification threshold T by using the ROC curve;
(6) Return S', T.

Algorithm 1: The algorithm flow of optimized Mahalanobis–Taguchi system.

Table 1: Description of the dataset.

Dataset name	Number of variables	Number of samples	Positive class	Negative class
Ionosphere	33	351	Good/225	Bad/126
Z-Alizadeh Sani	48	303	Normal/87	CAD/216
Parkinson dataset with replicated acoustic features	46	240	Healthy/120	PD/120
Breast Cancer Wisconsin (prognostic)	34	194	Recurrent/148	Nonrecurrent/46
Connectionist Bench (sonar, mines vs. rocks)	60	161	R/50	M/111

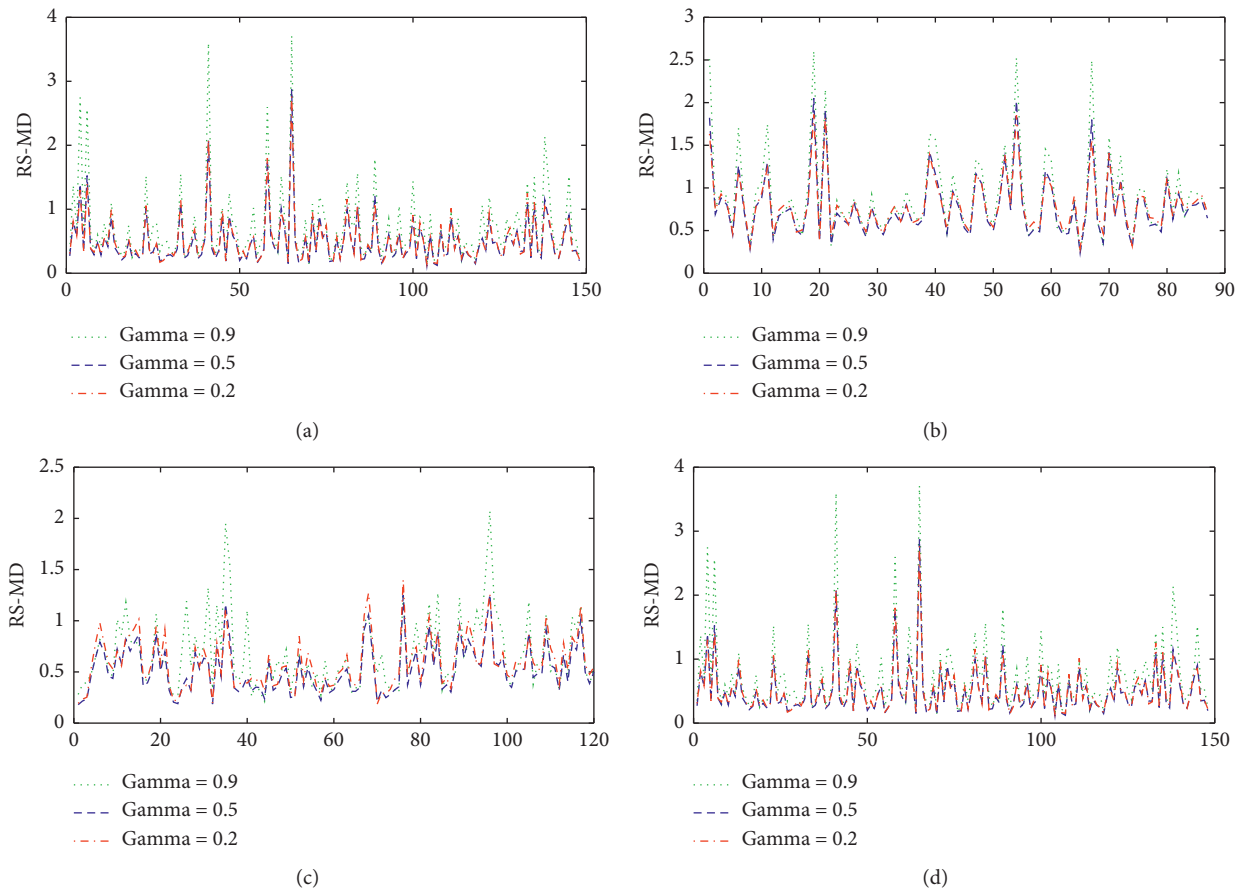

Figure 1: The distribution of RS-MD for normal observations. (a) Ionosphere. (b) Z-Alizadeh Sani. (c) Parkinson dataset with replicated acoustic features. (d) Breast Cancer Wisconsin (prognostic).

Table 2: The variance of the RS-MD.

Dataset name	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.5$	$\gamma = 0.9$
Ionosphere	0.1555	0.1182	0.0978	0.2463
Z-Alizadeh Sani	0.1127	0.1228	0.1483	0.2633
Parkinson dataset with replicated acoustic features	0.0650	0.0589	0.0552	0.1207
Breast Cancer Wisconsin (prognostic)	0.1419	0.1359	0.1455	0.3499

thereby resulting in an overfitting problem. The comparison of the variances shows that the comprehensive effect is better when parameter γ is 0.3. Therefore, we set parameter γ as 0.3 and smooth the eigenvalues less than 0.01. The calculation results of RS-MD and traditional MD are shown in Figure 2.

Figure 2 depicts the distributions of MD and RS-MD for the normal observations of each dataset. The distributions of between MD and RS-MD in the Z-Alizadeh Sani dataset are relatively close. The RS-MD is slightly smaller than the traditional MD, and the volatility is slightly reduced. In the other three datasets, the RS-MD is smaller than the traditional MD, and the volatility is significantly reduced. Therefore, the Mahalanobis metric based on the regularization and smoothing techniques is more robust than the traditional Mahalanobis metric.

However, when the number of normal observations is smaller than that of features, the calculated sample covariance matrix is singular, and the traditional Mahalanobis distance cannot be calculated. In order to verify the validity of the RS-MD under this condition, the Gram–Schmidt Mahalanobis distance (GS-MD) is compared with the RS-MD. Taking the Connectionist Bench (sonar, mines vs. rocks) dataset as an example, the calculation results are shown in Figure 3. It can be seen from Figure 3 that although GS-MD can calculate the Mahalanobis distance of each sample, the Mahalanobis distance of the samples in two classes almost overlaps, making it difficult to distinguish the samples in two classes effectively. When the RS-MD is used, there is a significant difference in two classes. The RS-MD can be used as an index to distinguish the samples. Therefore, the RS-MD can be used as a metric when the number of normal samples is less than that of features, and the discrimination of the different samples can be improved.

4.2. Comparative Analysis between the Two-Stage Feature Selection Method and the Feature Selection of Traditional MTS. To verify the validity of the two-stage feature selection method, the stability and classification accuracy of feature selection are analyzed in this section. The data of each dataset are divided into five folds, four of which are used as training data. In order to measure the stability of feature selection, the Jaccard coefficient is used to calculate the similarity of feature subsets.

The Jaccard coefficient is a common measure of similarity, which is used to measure the similarities among sample sets [54]. For any two sets A and B, the Jaccard coefficient is defined as follows:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \tag{30}$$

The mean of Jaccard coefficient in five experiments is taken as a measure of the stability of feature selection. The results are shown in Figure 4. Figure 4 presents the stability of the feature subset obtained by using two feature selection methods in each dataset, where mRMR-SNR represents the two-stage feature selection method and SNR represents the feature selection method of traditional MTS. Compared with the traditional MTS using the SNR to screen variables, the effect of the mRMR-SNR is better. This result shows that the two-stage feature selection method is beneficial to the improvement of the robustness of feature selection.

On the basis of the results of the two feature selection methods, the classification accuracy in each dataset after feature selection is calculated. Decision tree, SVM, and kNN are used as classifiers to measure the classification accuracy. Five-fold cross validation is used to calculate the classification accuracy for each dataset, and the results are shown in Figure 5.

Figure 5 presents the classification accuracy calculated by different classifiers after using two feature selection methods for each dataset. It can be seen that according to the feature subset obtained from mRMR-SNR in each dataset, the classification accuracy calculated is higher. Thus, the two-stage feature selection method is helpful to select the effective features in classifying.

5. Empirical Analysis

The formation and development of email offer great convenience to daily life. However, large numbers of spam cases also cause many problems for users and service providers. Therefore, how to obtain effective emails becomes a concern, and email filtering has gradually become an important way [55]. The purpose of email filtering is to distinguish regular messages from spam; this objective belongs to a typical two-class problem. Traditional classification algorithms often require a large number of labeled emails as training samples, but the collection and tagging of a large number of emails greatly increase the cost of consumption. Hence, improving email filtering performance under small sample conditions is an important research issue [56]. The MTS does not depend on the distribution type of data and can achieve classification prediction after reducing dimension. It is a practical pattern recognition and classification prediction method for multidimensional variables. Thus, we apply the optimized MTS to email filtering under small sample conditions.

5.1. Data Preprocessing. This section takes the Spambase dataset as an example provided in the UCI database. The dataset contains 4,601 email samples (2,788 regular emails and 1813 spam emails). The text content of each email is described by 56 different variables and 1 attribute variable. A total of 360 emails (190 regular emails and 170 spam emails)

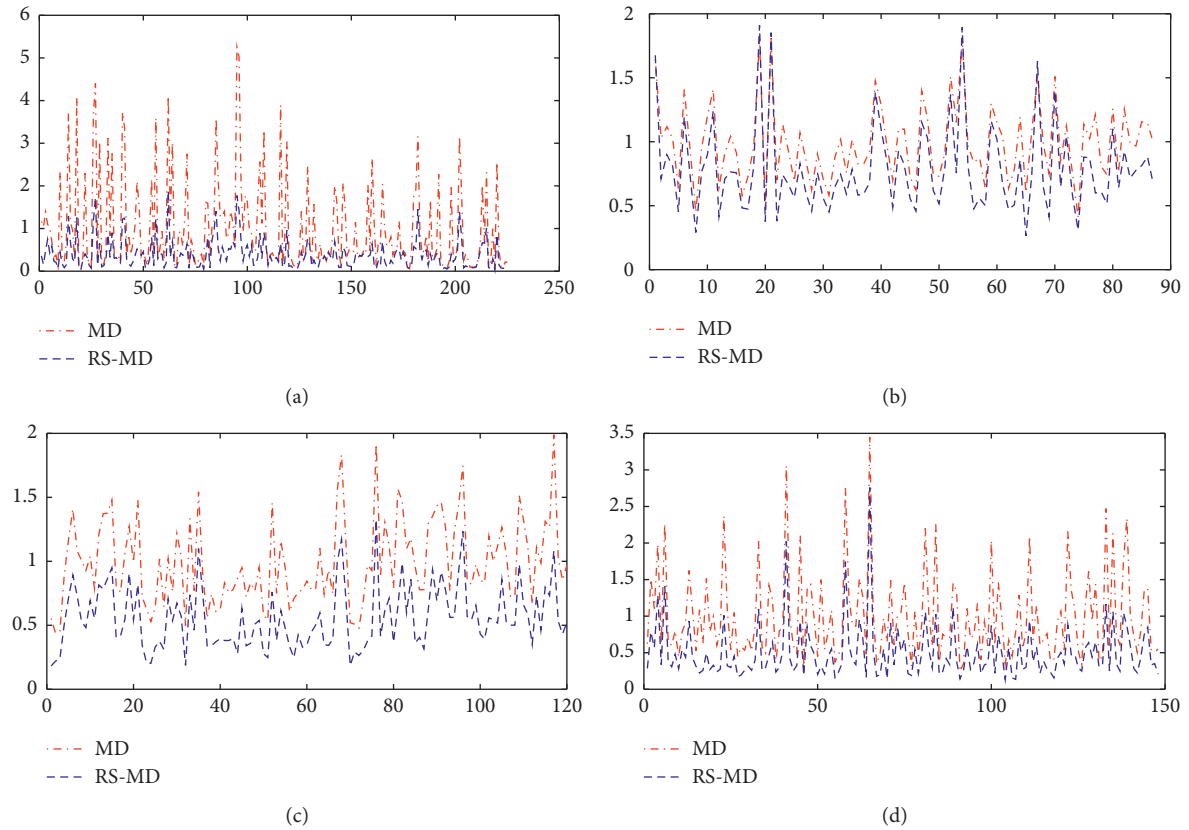

Figure 2: Comparison of the distributions between MD and RS-MD. (a) Ionosphere. (b) Z-Alizadeh Sani. (c) Parkinson dataset with replicated acoustic features. (d) Breast Cancer Wisconsin (prognostic).

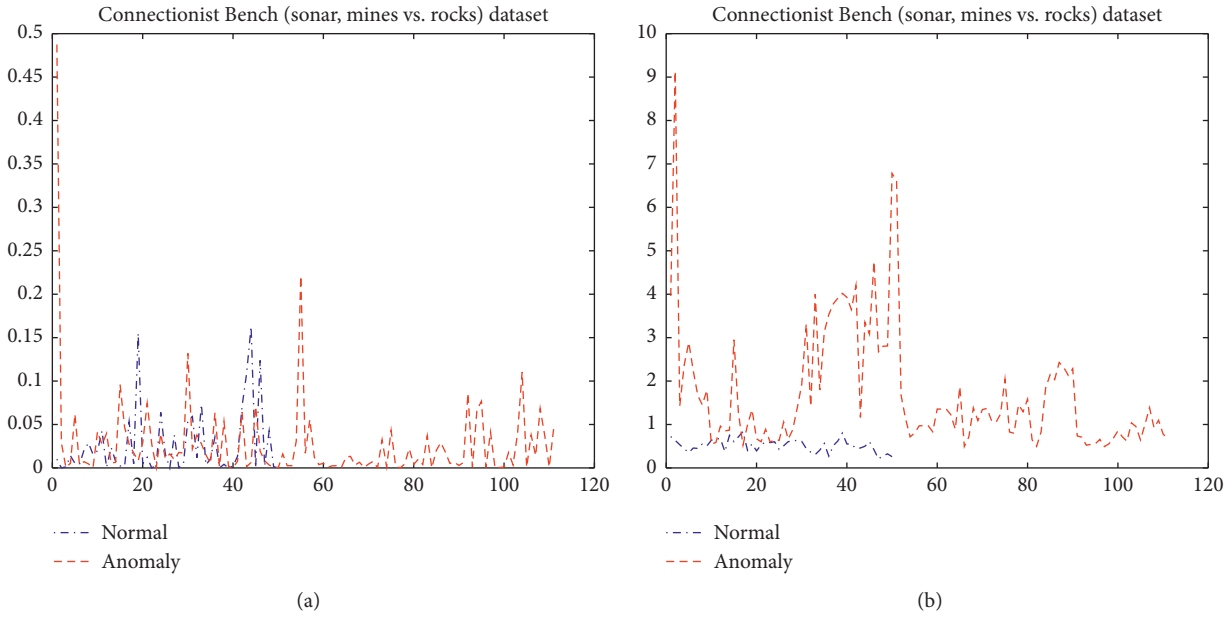

Figure 3: Comparison between RS-MD and GS-MD. (a) GS-MD. (b) RS-MD.

are randomly selected from the dataset to make up the training set, and the test set consists of 300 emails (160 regular emails and 140 spam emails). This is aimed at satisfying the requirements of high-dimensional small sample data and improving the efficiency of the algorithm.

5.2. Construction of Measurement Scale Based on Modified Mahalanobis Metric. In the training set, 190 regular emails are normal observations and 170 spam emails are abnormal observations. The RS-MD of each observation is calculated by equation (9). According to the results of the verification

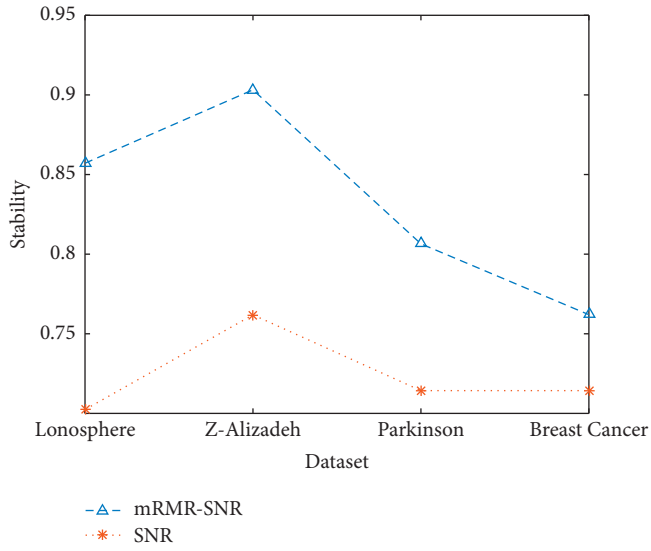

Figure 4: Stability of feature selection results for each dataset.

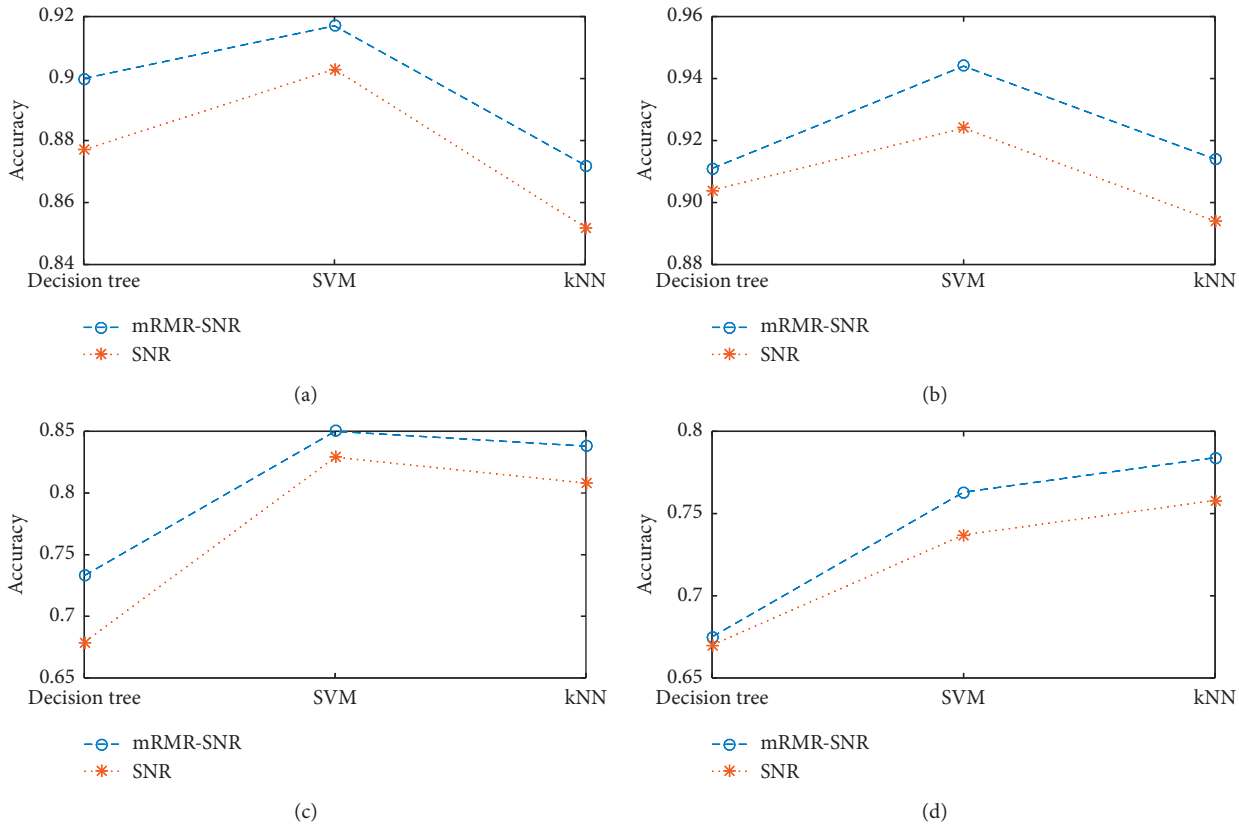

Figure 5: Classification accuracy after feature selection of each dataset. (a) Ionosphere. (b) Z-Alizadeh Sani. (c) Parkinson dataset with replicated acoustic features. (d) Breast Cancer Wisconsin (prognostic).

analysis, parameter γ is set as 0.3, and eigenvalues less than 0.01 are smoothed. According to the calculation results, the RS-MD of most of the abnormal observations is bigger than that of the normal observations, whereas the RS-MD of the normal observations are basically concentrated at approximately 1. Thus, the constructed reference space is effective.

5.3. Two-Stage Feature Selection Based on mRMR Algorithm and SNR. The original data consist of 56 variables, which are recorded as $X_1, X_2, \ldots, X_{56}$. The mRMR algorithm is first used to remove the noise variables and the redundant variables. The correlation is measured by mutual information, which is calculated by equations (25) and (26). The 31 features are retained, and their scores are shown in Table 3.

Table 3: Selected feature and its score by mRMR algorithm.

Feature	X_{18}	X_{46}	X_{28}	X_{52}	X_{40}	X_2	X_{26}	X_{21}	X_6	X_{43}	X_{37}
Score	0.000	−0.094	−0.019	−0.007	−0.014	−0.027	−0.026	−0.019	−0.016	−0.017	−0.030
Feature	X_{41}	X_{23}	X_{24}	X_{45}	X_{55}	X_{51}	X_{32}	X_{47}	X_{15}	X_{38}	X_{13}
Score	−0.022	−0.028	−0.027	−0.033	−0.041	−0.050	−0.052	−0.060	−0.057	−0.061	−0.066
Feature	X_{31}	X_{19}	X_{22}	X_{25}	X_{10}	X_{53}	X_{36}	X_{16}	X_{48}		
Score	−0.068	−0.063	−0.076	−0.072	−0.085	−0.092	−0.094	−0.096	−0.098		

Table 4: Signal-to-noise ratio values under each test.

Test	1	2	3	4	5	6	7	8	9	10	11
SNR	1.513	−1.271	−4.010	0.692	−7.809	−2.428	−3.104	−9.281	−5.210	−6.468	−5.422
Test	12	13	14	15	16	17	18	19	20	21	22
SNR	−5.690	−5.506	−4.236	−3.931	−2.749	−4.634	−4.227	−4.107	−6.069	−0.658	−2.976
Test	23	24	25	26	27	28	29	30	31	32	
SNR	−2.747	0.260	−3.678	−4.094	−4.759	−3.800	−4.712	−3.734	−5.076	−4.008	

An orthogonal table $L_{32}(2^{31})$ is selected on the basis of the 31 selected features. On the basis of the information of the orthogonal table, the RS-MD of each abnormal observation under different feature combinations is recalculated, and the SNR is calculated according to equation (30). The SNR values obtained from each test are shown in Table 4.

Combined with the orthogonal table and SNR, the mean of SNR of each feature at different levels is analyzed. The mean of SNR reflects the effect of the variable at different levels. When the mean of SNR at level 1 is greater than level 2, indicating that using this variable is more advantageous than not using it. That is, those variables are effective variables and are beneficial to classification. Conversely, when the mean of SNR at level 1 is lower than level 2, indicating that the effect of using this variable is lower than not using it. That is, those variables exert a slight influence on the classification and can be deleted. For valid variables, the difference of the mean of SNR at different levels reflects the significance of the variable. The greater the difference, the greater the contribution of the variable to the classification. Therefore, the reduced benchmark space is composed of these 15 variables $X_6, X_{10}, X_{15}, X_{16}, X_{18}, X_{19}, X_{22}, X_{23}, X_{24}, X_{36}, X_{38}, X_{51}, X_{52}, X_{53}, X_{55}$.

5.4. Threshold Calculation and Classification Prediction. The reference space is reconstructed in accordance with the 15 variables selected in the feature selection process, and the MD of each sample in the new reference space is calculated. The traditional MD is then used because the data after feature selection are no longer high-dimensional small sample data. Based on the calculated MD, the ROC curve is used to determine the threshold of the system. The result is shown in Figure 6. When the threshold is 1.9377, the classification accuracy of the training set reaches a maximum of 0.9194. The determined threshold is used to classify the test set, and the classification accuracy of the test set is 0.9067 eventually.

5.5. Comparison with Common Classification Methods for High-Dimensional Small Sample Data. For the classification problem of high-dimensional small sample data, a feature selection method, such as filter and embedded methods, is first used to filter the variables, and then a common machine learning algorithm is used to classify the dimension-reduced dataset. The relief method and the SVM-RFE method are the commonly used methods in the filter and embedded method, respectively. Therefore, this section first uses the relief and SVM-RFE methods to reduce the dimension of high-dimensional small sample data. Then, the decision tree, SVM, and kNN algorithm are used to classify the reduced-dimensional dataset.

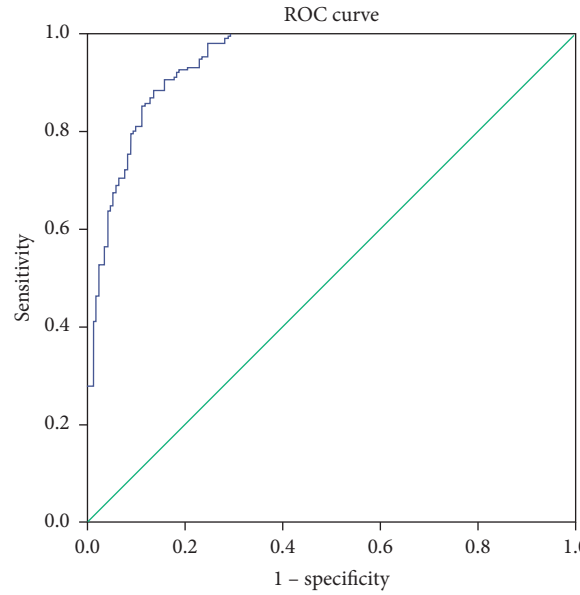

Figure 6: ROC curve corresponding to the Mahalanobis distance of the training set after feature selection.

Fifteen variables are selected by the relief or SVM-RFE method, and then the dimension-reduced dataset is classified by decision tree, SVM, and kNN algorithm. The results are compared with those of the optimized MTS. The comparison results are shown in Table 5.

As shown in Table 5, the optimized MTS has better classification effect than the classical MTS for training and test samples. This result shows that compared with the

Table 5: Comparison of results between optimized MTS and the classification methods for high-dimensional small sample data.

	Optimized MTS	Classical MTS	Decision tree	
			Relief	SVM-RFE
Number of features	15	20	15	15
Training set	0.9194	0.8722	0.8333	0.8444
Test set	0.9067	0.8633	0.8433	0.8533

	SVM		kNN	
	Relief	SVM-RFE	Relief	SVM-RFE
Number of features	15	15	15	15
Training set	0.8722	0.8944	0.8583	0.8750
Test set	0.8733	0.8967	0.8600	0.8867

classical MTS, the classification and prediction capability of the optimized MTS is better. That is, the optimized MTS is more suitable for small sample data.

After screening feature by relief and SVM-RFE methods, the classification effect of the SVM algorithm is better than that of the decision tree and kNN algorithm. This result shows that the SVM algorithm has the better classification performance under the condition of small samples. However, the classification effect of the three classifiers is lower than that of the optimized MTS. The optimized MTS has good dimension reduction and classification performance for high-dimensional small sample data. Moreover, the dimension reduction and classification prediction are separated in the commonly used classification methods for high-dimensional small sample data. By contrast, the optimized MTS can complete classification prediction after reducing variables, that is, solve the problem of dimension reduction and classification prediction at the same time. The optimized MTS thus maintains work efficiency to a certain extent.

6. Conclusion

This paper proposes the optimized MTS for high-dimensional small sample data. Aimed at the inverse matrix instability problem of the covariance matrix, a Mahalanobis metric based on regularization and smoothing techniques is proposed. Aimed at the feature selection problem, a two-stage feature selection algorithm based on the mRMR algorithm and SNR is proposed. Through the verification analysis of five datasets, the robustness of the modified Mahalanobis metric and the effectiveness of the two-stage feature selection method are verified. The optimized MTS is applied to email filtering problems under small sample conditions and achieves a good classification and dimension reduction effect. Simultaneously, compared with the classical MTS and the commonly used classification algorithms for high-dimensional small sample data, the optimized MTS performs better. Thus, the optimized MTS not only improves the generalization capability of the MTS but also provides a new approach for high-dimensional small sample data.

Data Availability

The calculation software used in this paper includes MATLAB 2016a, SPSS 22, and Minitab 17. The UCI database is available online at https://archive.ics.uci.edu/ml/datasets.php.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (71871174).

References

[1] H. Sakeran, N. A. Abu Osman, and M. S. Abdul Majid, "Gait classification using mahalanobis-taguchi system for health monitoring systems following anterior cruciate ligament reconstruction," *Applied Sciences*, vol. 9, no. 16, pp. 3306–3323, 2019.

[2] Y.-H. Hsiao, C.-T. Su, and P.-C. Fu, "Integrating MTS with bagging strategy for class imbalance problems," *International Journal of Machine Learning and Cybernetics*, vol. 2019, 2019.

[3] Y. P. Gu, L. S. Cheng, and Z. P. Chang, "Classification of imbalanced data based on MTS-CBPSO method: a case study of financial distress prediction," *Journal of Information Processing Systems*, vol. 15, no. 3, pp. 682–693, 2019.

[4] C. C. Wang and B. D. Wu, "Classification and prediction of wafer probe yield in DRAM manufacturing using Mahalanobis-Taguchi system and neural network," *South African Journal of Industrial Engineering*, vol. 30, no. 1, pp. 248–256, 2019.

[5] H. L. Lim, E.-H. Huh, D.-A. Huh, J.-R. Sohn, and K. W. Moon, "Priority setting for the management of chemicals using the globally harmonized system and multivariate analysis: use of the mahalanobis-taguchi system," *International Journal of Environmental Research and Public Health*, vol. 16, no. 17, pp. 3119–3130, 2019.

[6] N. Wang, Z. P. Wang, L. Jia et al., "Adaptive multiclass Mahalanobis-Taguchi system for bearing fault diagnosis under variable conditions," *Sensors*, vol. 19, no. 1, pp. 26–41, 2019.

[7] J. Yuan and X. Luo, "Regional energy security performance evaluation in China using MTGS and SPA-TOPSIS," *Science of The Total Environment*, vol. 696, no. 8, p. 133817, 2019.

[8] M. Rizal, J. A. Ghani, M. Z. Nuawi, and C. H. C. Haron, "Cutting tool wear classification and detection using multi-sensor signals and Mahalanobis-Taguchi System," *Wear*, vol. 376-377, no. 1, pp. 1759–1765, 2017.

[9] S. Sikder, S. C. Panja, and I. Mukherjee, "An integrated approach for multivariate statistical process control using Mahalanobis-Taguchi System and Andrews function," *International Journal of Quality & Reliability Management*, vol. 34, no. 8, pp. 1186–1208, 2017.

[10] P. Chi-Feng, H. Li-Hsing, T. Sang-Bing et al., "Applying the Mahalanobis-Taguchi system to improve tablet PC production processes," *Sustainability*, vol. 9, no. 9, pp. 1557–1573, 2017.

[11] E. B. Mahmoud, "Modified Mahalanobis-Taguchi system for imbalance data classification," *Computational Intelligence and Neuroscience*, vol. 2017, pp. 5874896–5874910, 2017.

[12] A. C. M. Miguel, S. M. Italo, and S. Gilberto, "Comparing principal component analysis and Mahalanobis-Taguchi system to detect unbalance in a centrifugal compressor in a floating production storage & offloading," *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, vol. 6, pp. 41–85, 2019.

[13] F. Samadi, G. Akbarizadeh, and H. Kaabi, "Change detection in SAR images using deep belief network: a new training approach based on morphological images," *IET Image Processing*, vol. 13, no. 12, pp. 2255–2264, 2019.

[14] F. Sharifzadeh, G. Akbarizadeh, and Y. Seifi Kavian, "Ship classification in SAR images using a new hybrid CNN-MLP classifier," *Journal of the Indian Society of Remote Sensing*, vol. 47, no. 4, pp. 551–562, 2019.

[15] A. E. Moghaddam, G. Akbarizadeh, and H. Kaabi, "Automatic detection and segmentation of blood vessels and pulmonary nodules based on a line tracking method and generalized linear regression model," *Signal, Image and Video Processing*, vol. 13, no. 3, pp. 457–464, 2019.

[16] M. Norouzi, G. Akbarizadeh, and F. Eftekhar, "A hybrid feature extraction method for SAR image registration," *Signal, Image and Video Processing*, vol. 12, no. 8, pp. 1559–1566, 2018.

[17] G. Akbarizadeh and M. Rahmani, "Efficient combination of texture and color features in a new spectral clustering method for PolSAR image segmentation," *National Academy Science Letters*, vol. 40, no. 2, pp. 117–120, 2017.

[18] G. Akbarizadeh, "A new statistical-based kurtosis wavelet energy feature for texture recognition of SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 11, pp. 4358–4368, 2012.

[19] A. Raeisi, G. Akbarizadeh, and A. Mahmoudi, "Combined method of an efficient cuckoo search algorithm and non-negative matrix factorization of different zernike moment features for discrimination between oil spills and lookalikes in SAR images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 4193–4205, 2018.

[20] M. Modava, G. Akbarizadeh, and M. Soroosh, "Integration of spectral histogram and level set for coastline detection in SAR images," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 2, pp. 810–819, 2019.

[21] G. Akbarizadeh, M. Modava, and M. Soroosh, "Hierarchical coastline detection in SAR images based on spectral-textural features and global-local information," *IET Radar, Sonar & Navigation*, vol. 10, 2019.

[22] X. Xiao and H. Duan, "A new grey model for traffic flow mechanics," *Engineering Applications of Artificial Intelligence*, vol. 88, p. 103350, 2020.

[23] A. Karaivanova, S. Ivanovska, and T. Gurov, "Monte Carlo method for density reconstruction based on insufficient data," *Procedia Computer Science*, vol. 51, no. 10, pp. 1782–1790, 2015.

[24] H.-F. Gong, Z.-S. Chen, Q.-X. Zhu, and Y.-L. He, "A Monte Carlo and PSO based virtual sample generation method for enhancing the energy prediction and energy optimization on small data problem: an empirical study of petrochemical industries," *Applied Energy*, vol. 197, no. 7, pp. 405–415, 2017.

[25] D. P. Tao, Y. N. Guo, M. L. Song et al., "Person re-identification by dual-regularized KISS metric learning," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2726–2738, 2016.

[26] V. E. Liong, Y. X. Ge, and J. W. Lu, *Discriminative Regularized Metric Learning for Person re-identification*, IEEE, in *Proceedings of the International Conference on Biometrics*, September 2015.

[27] S. Espezua, E. Villanueva, C. D. Maciel, and A. Carvalho, "A Projection Pursuit framework for supervised dimension reduction of high dimensional small sample datasets," *Neurocomputing*, vol. 149, no. 16, pp. 767–776, 2015.

[28] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances in Bioinformatics*, vol. 2015, no. 5, pp. 1–13, 2015.

[29] S. Kamyab and M. Eftekhari, "Feature selection using multimodal optimization techniques," *Neurocomputing*, vol. 171, no. 3, pp. 586–597, 2016.

[30] W. W. B. Goh and L. Wong, "Evaluating feature-selection stability in next-generation proteomics," *Journal of Bioinformatics and Computational Biology*, vol. 14, no. 5, pp. 29–51, 2016.

[31] W. Du, Z. B. Cao, T. C. Song et al., "A feature selection method based on multiple kernel learning with expression profiles of different types," *BioData Mining*, vol. 10, no. 1, pp. 4–19, 2017.

[32] E. A. Cudney and K. M. Ragsdell, *Forecasting Using the Mahalanobis-Taguchi System in the Presence of Collinearity*, SAE World Congress & Exhibition, Detroit, MI, USA, 2006.

[33] E. A. Cudney, K. Paryani, and K. M. Ragsdell, "Identifying useful variables for vehicle braking using the adjoint matrix approach to the Mahalanobis-Taguchi system," *International Journal of Industrial & Systems Engineering*, vol. 1, no. 4, pp. 281–292, 2008.

[34] C.-T. Su and Y.-H. Hsiao, "An evaluation of the robustness of MTS for imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 10, pp. 1321–1332, 2007.

[35] P. Shakya, M. S. Kulkarni, and A. K. Darpe, "Bearing diagnosis based on Mahalanobis-Taguchi-Gram-Schmidt method," *Journal of Sound and Vibration*, vol. 337, no. 43, pp. 342–362, 2015.

[36] Y. J. Han, W. He, and F. Song Guo, "Research on the related problems of Mahalanobis-Taguchi system in multi-dimensional system optimization," *Industrial Engineering Journal*, vol. 15, no. 2, pp. 71–77, 2012.

[37] Z. P. Chang, L. S. Cheng, and J. S. Liu, "Interval number multi-attribute decision making method based on Mahalanobis-Taguchi system and TOPSIS," *Systems Engineering Theory & Practice*, vol. 34, no. 1, pp. 168–175, 2014.

[38] J. B. Tao and L. S. Cheng, "Application of ridge Mahalanobis-Taguchi system in complex colinear data based on ridge estimation," *Journal of Mathematics in Practice and Theory*, vol. 46, no. 4, pp. 109–116, 2016.

[39] B. Abraham and A. M. Variyath, "Discussion," *Technometrics*, vol. 45, no. 1, pp. 22–24, 2003.

[40] E. Reséndiz, L. A. Moncayo-Martínez, and G. Solís, "Binary ant colony optimization applied to variable screening in the

Mahalanobis-Taguchi System," *Expert Systems with Applications*, vol. 40, no. 2, pp. 634–637, 2013.

[41] A. S. Iquebal, A. Pal, D. Ceglarek, and M. K. Tiwari, "Enhancement of Mahalanobis-Taguchi system via rough sets based feature selection," *Expert Systems with Applications*, vol. 41, no. 17, pp. 8003–8015, 2014.

[42] E. O. Reséndiz-Flores, J. A. Navarro-Acosta, and A. Hernández-Martínez, "Optimal feature selection in industrial foam injection processes using hybrid binary particle swarm optimization and gravitational search algorithm in the Mahalanobis-Taguchi system," *Soft Computing*, vol. 23, no. 6, pp. 1–9, 2019.

[43] Y. I. Reyes-Carlos, C. G. Mota-Gutiérrez, and E. O. Reséndiz-Flores, "Optimal variable screening in automobile motorhead machining process using metaheuristic approaches in the Mahalanobis-Taguchi System," *The International Journal of Advanced Manufacturing Technology*, vol. 95, no. 9–12, pp. 3589–3597, 2018.

[44] E. B. Mahmoud, "A novel approach for classifying imbalance welding data: Mahalanobis genetic algorithm (MGA)," *The International Journal of Advanced Manufacturing Technology*, vol. 77, no. 4, pp. 407–425, 2015.

[45] J. L. Niu and L. S. Cheng, "An unbalanced data classification method based on improved Mahalanobis-Taguchi system," *Journal of Industrial Engineering and Engineering Management*, vol. 26, no. 2, pp. 85–93, 2012.

[46] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.

[47] F. Kimura, K. Takashina, S. Tsuruoka, and Y. Miyake, "Modified quadratic discriminant functions and the application to Chinese character recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 1, pp. 149–153, 1987.

[48] D. P. Tao, L. W. Jin, Y. F. Wang et al., "Person re-identification by minimum classification error-based KISS metric learning [J]," *IEEE Transactions on Cybernetics*, vol. 2, no. 45, pp. 242–252, 2015.

[49] X. Zhang, Z. Song, D. Li, W. Zhang, Z. Zhao, and Y. Chen, "Fault diagnosis for reducer via improved LMD and SVM-RFE-MRMR," *Shock and Vibration*, vol. 2018, no. 7, pp. 1–13, 2018.

[50] L. Huang, Z. J. Xiang, and H. Chu, "Remote sensing image classification algorithm based on mRMR selection and IFCM clustering," *Bulletin of Surveying and Mapping*, vol. 4, pp. 32–37, 2019.

[51] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[52] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, pp. 1–16, 2004.

[53] X. Xu, K. Zhang, and W. J. Wang, "A feature selection method for small sample data," *Journal of Computer Research and Development*, vol. 55, no. 10, pp. 229–238, 2018.

[54] B. K. Samanthula and W. Jiang, "Secure multiset intersection cardinality and its application to jaccard coefficient," *IEEE Transactions on Dependable and Secure Computing*, vol. 13, no. 5, pp. 591–604, 2015.

[55] L. Shu, K. Mcisaac, G. R. Osinski, and R. Francis, "Unsupervised feature learning for autonomous rock image classification," *Computers & Geosciences*, vol. 106, pp. 10–17, 2017.

[56] J. Z. Pan, X. Zhou, G. Q. Wu et al., "Spam filtering method based on small sample learning," *Computer Engineering*, vol. 36, no. 21, pp. 245–247, 2010.