A Novel Effective Dimensionality Reduction Algorithm for Water Chiller Fault Data

Zhuozheng Wang[1,2], Yingjie Dong[1,2,*], Wei Liu[1,2]
[1] Faculty of Information Technology, Beijing University of Technology, Beijing, China
[2] Intelligent Signal Processing Laboratory, Beijing University of Technology, Beijing, China
E-mail: wangzhuozheng@bjut.edu.cn
[*] E-mail: yingjied@emails.bjut.edu.cn

Abstract—The reliability of chiller is very important for the safe operation of refrigeration system. In order to solve the problem that the traditional linear discriminant analysis (LDA) based on L_2 norm is sensitive to outliers, this paper introduced a novel dimensionality reduction algorithm for chiller fault data set – RSLDA. Firstly, $L_{2,1}$ norm is used to extract the most discriminant features adaptively and eliminate the redundant features instead of L_2 norm. Secondly, an orthogonal matrix and a sparse matrix are introduced to ensure the extracted features contain the main energy of the raw features. In addition, the recognition rate of the nearest classifier is defined as the performance criteria to evaluate the effectiveness of dimensionality reduction. Finally, the reliability of algorithm was verified by experiences compared with other algorithms. Experimental results revealed that RSLDA not only improves robustness but also has a good performance in the Small Sample Size problem (SSS) of fault classification.

I. INTRODUCTION

As the key equipment of refrigeration system, chiller is mainly used to provide cold source to maintain stable temperature for IDC (Internet Data Center) room [1]. The occurrence of the chiller fault not only reduces the performance of the refrigeration system but also shortens the service life cycle of the equipment. Especially, the loss of data stored in the computers leads to serious or even irreparable economic losses. Thus, discovering potential faults and hidden danger in the process of operation of chillers in time can guarantee the reliability and security of IDC. Refrigeration system of IDC is shown as Fig. 1.

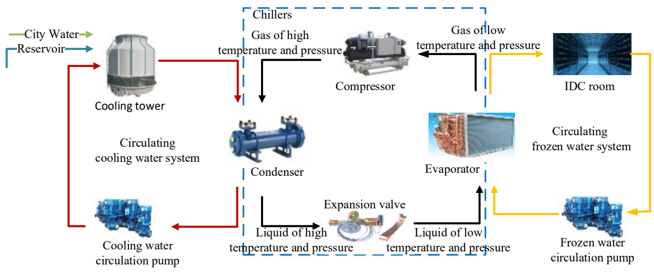


Fig. 1 Refrigeration system of Internet Data Center.

The real-time running data of the equipment are collected by setting up multiple sensors, and the quantitative indicators from the time domain, frequency domain and other aspects that can fully comprehensively reflect the fault state are selected [2][3]. However, sensor faults may lead to outliers in the collected data. Furthermore, traditional linear discriminant analysis requires sufficient train data to avoid the Small Sample Size problem (SSS) [4], which makes it difficult to deal with the small-scale data with high- dimension [5]. In view of the above problems, it is particularly important to find a dimensionality reduction method that can effectively eliminate redundant information, improve the robustness of outliers and overcome the SSS problem to improve the accuracy of fault diagnosis.

The purpose of dimensionality reduction is to find a most discriminating low-dimensional subspace to replace the raw high-dimensional space and ensure the minimum loss of information in the process. Dimensionality reduction is the key to solve the problem of " curse of dimensionality " in the field of pattern recognition and machine learning, because it not only reduces the complexity of data, but also improves the efficiency and accuracy of the classification algorithm. Dimensionality reduction techniques can be divided into feature selection approaches and feature extraction approaches. Feature selection approaches aim to select the most important or representative feature subset from the raw feature set, and all the selected features can be found in the raw feature set. Feature extraction approaches transform raw features into a new feature space with lower dimensionality and the newly constructed features are usually non-linear combinations of the raw features. Dimensionality reduction can be divided into unsupervised and supervised dimensionality reduction according to whether the samples are labelled or not. Principal component analysis (PCA) [6] is the most classical unsupervised dimensionality reduction method. Its goal is to learn a set of projection vectors so that the variance of given data in low dimensional feature space is maximized. However, the most serious disadvantage of PCA is that the category information in the training set is not taken into account in classification tasks. Linear Discriminant Analysis (LDA) [7] is the most popular supervised dimensionality reduction method, which is designed to find a projection matrix that maximizing between-class covariance and minimizing within-class covariance. That is, LDA aims to find the vector that maximizes the Fisher criterion function [8] as the optimal projection direction.

The objective function of traditional LDA is calculated based on L_2 norm. The projection direction obtained is unstable when the training data set includes outliers. Because the square operation in the L_2 norm amplifies the influence of outliers [9]. Although the influence of outliers can be mitigated by assigning smaller weights to outliers [10], it is difficult to determine the optimal value of weighted parameters. Compared with L_2 norm, L_1 norm is more robust to outliers because the absolute value operation reduces the influence of outliers. Based on this, many related algorithms have been proposed. A method of maximizing L_1 norm in feature space is proposed by Nojun Kwak [11], which not only achieves the robustness to outliers, but also has the invariance of rotation. However, it is easy to fall into the problem of local optimal solution because the obtained projection is calculated one by one through greedy search strategy. To solve this problem, Feiping Nie et al. [12] proposed a principal component analysis algorithm based on non-greedy strategy, which made all projection directions optimized. L1-LDA algorithm theoretically overcomes the problem of singular solution of LDA intraclass scattering matrix based on L_1 norm and achieves robustness to outliers [13]. A new linear discriminant analysis method, LDA-L1, which maximizes the ratio of inter-class dispersion to intra-class dispersion by using L_1 norm instead of L_2 norm, and solves a series of local optimal projection vectors through iteration algorithm, was proposed in dimensionality reduction [14]. Lu G F et al. [15] proposed a new linear discriminant analysis method based on sparse L_1 norm, SLDA-L1, which finds the optimal projection vector that maximizes the inter-class scattering matrix and minimizes the intra-class scattering matrix, and regularizes the base vector by using the elastic network, where the L_1 norm is used for both robust modeling and sparse modeling, are proposed. In addition, this paper also proposed an effective iterative algorithm for SLDA-L1. A new robust feature selection method with joint $L_{2,1}$ norm minimization in both loss function and regularization are proposed [16]. In addition, an effective algorithm to minimize the norm of $L_{2,1}$ is proposed and the effectiveness of the algorithm is proved. Reference [17] proposed a new effective robust classification algorithm based on $L_{2,1}$ norm and gave the objective of regularized $L_{2,1}$ norm minimization model. The advantages of this algorithm are as follows: firstly, the loss function is robust to the outliers in the sample; secondly, the regularization phase can select the samples of the feature group in the whole training set, which has reasonable sparsity. Reference [18] proposed a robust linear discriminant dimension reduction method based on $L_{2,1}$ norm, RLDA, which takes the class weighted mean points instead of the class mean points as the class center points to suppress the influence of outliers. In addition, an efficient iterative algorithm for solving optimization problems is proposed and its convergence is proved. Reference [19] put LDA on top of a deep neural network and proposed the deep linear discriminant analysis (DeepLDA) for classification problems. DeepLDA has good performance on the large-scale image datasets. But, this paper mainly focused on the conventional linear discriminant analysis methods on the dimensionality reduction tasks with small scale water chiller fault data.

To overcome the SSS problem, numerous extensions of LDA have been proposed, For example, Regularized LDA (RLDA) [20], Orthogonal LDA (OLDA) [21] and so on.

Inspired by Ref. [22], RSLDA is introduced to the field of dimensionality reduction for fault data in this paper. Based on $L_{2,1}$, an orthogonal matrix and a sparse matrix are employed to preserve the main energy of the raw feature. Different from the traditional LDA, the adopted method is more flexible in dimensionality selection and more robust to outliers. The main contributions of this paper can be summarized as follows:

(1) A novel robust dimensionality reduction algorithm via joint $L_{2,1}$ norm with LDA is introduced in fault diagnosis to suppress the influence of outliers for improving the accuracy of fault diagnosis.
(2) The method has a good performance in the small-scale data with high-dimensionality.
(3) The method is more flexible in dimensionality selection.

The remainder of this paper is arranged as follows: In Section II, the traditional LDA algorithm and K-Nearest Neighborhood (KNN) are briefly reviewed; Section III introduces the RSLDA algorithm in detail; Data Preparation and Experimental Process are introduced in Section IV; Section V presents our experimental results and makes corresponding analysis; The conclusion is presented in section VI.

II. Related Work

A. Norm Specification

This paper use lowercase bold letters and uppercase bold letters to represent vectors and matrices, respectively. Assuming there is a vector $\boldsymbol{a} \in \boldsymbol{R}^n$, the L_p norm is calculated as:

$$\|\boldsymbol{a}\|_p = \left(\sum_{i=1}^{n} |a_i|^p\right)^{1/p} . \tag{1}$$

Where $p \geq 1$ is a real number [17]. For a matrix $\boldsymbol{A} = \{a_{ij}\} \in \boldsymbol{R}^{m\times n}$, where $\boldsymbol{a^i}$ and $\boldsymbol{a_j}$ represent the row and column vectors of matrix $\boldsymbol{A}$ respectively. The L_p norm of matrix $\boldsymbol{A}$ is calculated as:

$$\|\boldsymbol{A}\|_p = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^p\right)^{1/p} . \tag{2}$$

When $p = 2$, the L_F norm of the matrix $\boldsymbol{A}$ is calculated as:

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} {a_{ij}}^2} . \tag{3}$$

The $L_{2,1}$ norm of the matrix $\boldsymbol{A}$ is calculated as:

$$\|\boldsymbol{A}\|_{2,1} = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} {a_{ij}}^2} = \sum_{j=1}^{n} \|\boldsymbol{a}^i\|_2 . \tag{4}$$

B. Mathematical Preparations

Hermitian Matrix: Transposition is equal to the matrix itself, which means that every element of the ith row and the jth column of the matrix is equal to the conjugate of the element of the jth row and the ith column:

$$\boldsymbol{X}^T = \boldsymbol{X} . \tag{5}$$

Generalized Rayleigh Entropy: If $\boldsymbol{x}$ is a non-zero vector and $\boldsymbol{A}$ and $\boldsymbol{B}$ are Hermitian matrices, then

$$\mathrm{R}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{x}) = \frac{\boldsymbol{x^T A x}}{\boldsymbol{x^T B x}} . \tag{6}$$

Equation (6) denotes the generalized Rayleigh entropy of $\boldsymbol{A}$ and $\boldsymbol{B}$. The maximum eigenvalue of $\boldsymbol{B}^{-1}\boldsymbol{A}$ is its maximum.

Singular value decomposition of matrix: Any real matrix $\boldsymbol{M}$ can be decomposed into the product of the three matrices:

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\boldsymbol{T}}. \tag{7}$$

Where $\boldsymbol{U}$ and $\boldsymbol{V}$ are both orthogonal matrices. The column vector of $\boldsymbol{U}$ is the eigenvector of $\boldsymbol{M}\boldsymbol{M}^{\boldsymbol{T}}$ and the column $\boldsymbol{V}$ is the eigenvector of $\boldsymbol{M}^{\boldsymbol{T}}\boldsymbol{M}$.

C. Traditional Linear Discriminant Analysis (LDA)

Given a high-dimensional dataset $\boldsymbol{X}$=[$x_1, x_2, \ldots\ldots, x_N$] $\in \boldsymbol{R}^{d\times n}$ where *d* and *n* represent the dimension and number of the raw feature, respectively. The goal of LDA is learning a linear projection matrix $\boldsymbol{W} \in \boldsymbol{R}^{d\times m}$ $(m \le d)$ which makes the projection points from the same class as close as possible and the projection points from different classes as far as possible. After mapping the raw high-dimensional feature space $\boldsymbol{x_i} \in \boldsymbol{R}^d$ into the low-dimensional space $\boldsymbol{y_j} \in \boldsymbol{R}^m$. The projection formula is shown as follows,

$$\boldsymbol{y_j} = \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{x_i} . \tag{8}$$

Fig. 2 illustrates that the projection effect of the left is superior to that of right.

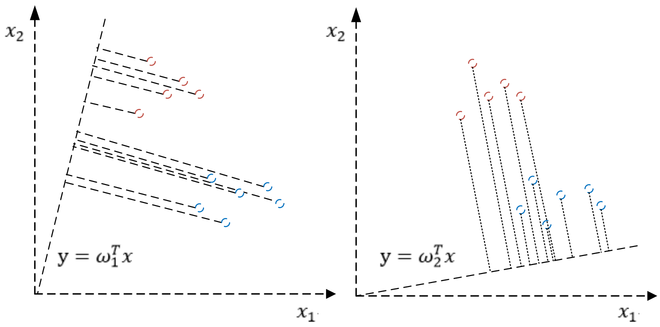


Fig. 2 The projection effect with different projection directions.

Assuming there are *c* classes in dataset and each class has n_i samples, the total number of samples is $n = \sum n_i$. x_j^i denotes the jth training sample of the ith class where *i=1,2,……c, j=1,2,……*n_i. According to the Fisher criterion, LDA needs to maximize the distance of projection points from different class L_b and minimize the distance of projection points from the same class L_w,

$$L_b = \sum_{i=1}^{c} n_i \|\overline{y_i} - \overline{y}\|_2^2 \ . \tag{9}$$

$$L_w = \sum_{i=1}^{c}\sum_{j=1}^{n_i} \left\|y_j^i - \overline{y_i}\right\|_2^2 \ . \tag{10}$$

where $\overline{y_i}$ and $\overline{y}$ denote the class center of the ith class sample and the class center of all samples after projection, respectively, y_j^i denotes the projection point of x_j^i. According $\boldsymbol{y_j} = \boldsymbol{W}^T\boldsymbol{x_i}$, the objective function of traditional LDA can be obtained,

$$J(\boldsymbol{W}) = \frac{\sum_{i=1}^{c} n_i\,(\overline{y_i} - \overline{y})((\overline{y_i} - \overline{y})^T}{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(y_j^i - \overline{y_i})\,(y_j^i - \overline{y_i})^T}$$

$$= \frac{\sum_{i=1}^{c} n_i \boldsymbol{W}^T\,(\overline{x_i} - \overline{x})((\overline{x_i} - \overline{x})^T\boldsymbol{W}}{\sum_{i=1}^{c}\sum_{j=1}^{n_i}\boldsymbol{W}^T\left(x_j^i - \overline{x_i}\right)\left(x_j^i - \overline{x_i}\right)^T\boldsymbol{W}}$$

$$= \frac{\boldsymbol{W}^T S_b \boldsymbol{W}}{\boldsymbol{W}^T S_w \boldsymbol{W}} \ . \tag{11}$$

where T denotes the transposition operation, $\bar{x} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j}^{n_i} x_j^i$ denotes the mean feature of all samples and $\overline{x_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_j^i$ denotes the mean feature of the ith class. S_b and S_w denote the inter-class divergence matrix and intra-class divergence matrix of the feature space respectively. The calculation formulas are as follows:

$$S_b = \sum_{i=1}^{c} n_i\,(\overline{x_i} - \bar{x})((\overline{x_i} - \bar{x})^T \ . \tag{12}$$

$$S_w = \sum_{i=1}^{c}\sum_{j=1}^{n_i}(x_j^i - \overline{x_i})\,(x_j^i - \overline{x_i})^T \ . \tag{13}$$

According to the Fisher criterion, the formula of the optimal projection matrix is obtained as follows:

$$\boldsymbol{W}_{opt} = \arg\max_{\boldsymbol{W}} Tr(J(\boldsymbol{W})) = \arg\max_{\boldsymbol{W}} \frac{Tr(\boldsymbol{W}^T S_b \boldsymbol{W})}{Tr(\boldsymbol{W}^T S_w \boldsymbol{W})} \ . \tag{14}$$

Where *Tr(.)* denotes the trace of matrix. The transformation matrix $\boldsymbol{W}$ is composed of eigenvectors corresponding to the first m eigenvalues of ${S_w}^{-1}S_b$ according to the generalized Rayleigh entropy. Generally, the optimal solution of (14) is equivalent to solving the following problem,

$$\boldsymbol{W}_{opt} = \arg\min_{\boldsymbol{W}^T\boldsymbol{W}=\boldsymbol{I}} Tr(\boldsymbol{W}^T(S_w - \lambda S_b)\boldsymbol{W}). \tag{15}$$

where λ is a small positive constant to balance the importance of the inter-class divergence matrix S_b and intra-class divergence matrix S_w. Since only $(c-1)$ of between-class scatter matrix is independent, the highest dimension of LDA to reduce is $(c-1)$ [23].

D. K- Nearest Neighborhood (KNN)

Nonparametric classification methods are a type of estimation methods that bypass probability and go directly to posterior probability estimation rather than determine the posterior function by parameter estimation [24]. As one of the simplest algorithms of nonparametric classification methods, the K-Nearest Neighborhood (KNN) needs reference data points for both classes. Secondly, each test sample is attributed the same class label as the label of the majority of its K nearest (reference) neighbors. Third, the Euclidean distance between the test point and all the reference points are calculated to find *K* nearest neighbors, and finally take the reference points corresponding to the *k* smallest Euclidean distances after sorting the obtained distances in ascending order [24].

III. ROBUST SPARSE LINEAR DISCRIMINANT ANALYSIS

A. Motivation and Problem Formulation

L_1 norm and $L_{2,1}$ norm have stronger robustness than L_2 norm [22]. Based on this, a more robustness projection matrix obtained by employed this constraint of $L_{2,1}$ norm is as follows:

$$\min_{\boldsymbol{W}} Tr(\boldsymbol{W}^T(S_w - \lambda S_b)\boldsymbol{W}) + \lambda_1\|\boldsymbol{W}\|_{2,1},$$

$$\text{subject to } \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}. \tag{16}$$

where $\mathbf{W} \in \boldsymbol{R}^{d\times m}$ $(m \le d)$ is the projection matrix, λ_1 is a trade-off parameter. The variants of the PCA constraint are introduced into the objective function (16) to preserve the discriminative information:

$$\min_{\boldsymbol{W},\boldsymbol{P}} Tr(\boldsymbol{W}^T(S_w - \lambda S_b)\boldsymbol{W}) + \lambda_1\|\boldsymbol{W}\|_{2,1} \ ,$$

$$\text{subject to } \boldsymbol{X} = \boldsymbol{P}\boldsymbol{W}^T\mathbf{X}, \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}. \quad (17)$$

where $\boldsymbol{P} \in \boldsymbol{R}^{d \times m}$ is an orthogonal matrix. By introducing the constraint, RSLDA can reserve the main energy of the raw feature set after dimensionality reduction. Acquired data set often includes outliers in real application, so a spare term is imposed to reduce the effect of outliers. Therefore the objective function can be rewritten as follows:

$$\min_{\boldsymbol{W},\boldsymbol{P},\boldsymbol{E}} Tr(\boldsymbol{W}^T(S_w - \lambda S_b)\boldsymbol{W}) + \lambda_1\|\boldsymbol{W}\|_{2,1} + \lambda_2\|\boldsymbol{E}\|_1,$$
$$\text{subject to } \boldsymbol{X} = \boldsymbol{P}\boldsymbol{W}^T\boldsymbol{X} + \boldsymbol{E}, \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}. \quad (18)$$

where λ_2 denotes a trade-off parameter, $\boldsymbol{E}$ denotes error and is used to fit outliers.

B. *An Efficient Algorithm to Solve the Problem*

Alternating Direction Method of Multipliers (ADMM)[18] is introduced to solve the optimization problem of (18). The minimization problem and constraints in (18) are transformed into augmented Lagrangian functions. As shown in (19):

$$\begin{aligned} L(\boldsymbol{P},\boldsymbol{W},\boldsymbol{E},\boldsymbol{Y}) &= Tr(\boldsymbol{W}^T(S_w - \lambda S_b)\boldsymbol{W}) + \lambda_1\|\boldsymbol{W}\|_{2,1} \\ &\quad + \lambda_2\|\boldsymbol{E}\|_1 \quad + \langle \boldsymbol{Y}, \boldsymbol{X} - \mathbf{P}\boldsymbol{W}^T\mathbf{X} - \boldsymbol{E}\rangle \\ &\quad + \frac{\beta}{2}\|\boldsymbol{X} - \mathbf{P}\boldsymbol{W}^T\mathbf{X} - \boldsymbol{E}\|_F^2 \\ &= Tr(\boldsymbol{W}^T(S_w - \lambda S_b)\boldsymbol{W}) + \lambda_1\|\boldsymbol{W}\|_{2,1} \\ &\quad + \lambda_2\|\boldsymbol{E}\|_1 + \boldsymbol{Y}^T(\boldsymbol{X} - \boldsymbol{P}\boldsymbol{W}^T\boldsymbol{X} - \boldsymbol{E}) \\ &\quad + \frac{\beta}{2}\|\boldsymbol{X} - \boldsymbol{P}\boldsymbol{W}^T\boldsymbol{X} - \boldsymbol{E}\|_F^2 \\ &= Tr(\boldsymbol{W}^T(S_w - \lambda S_b)\boldsymbol{W}) + \lambda_1\|\boldsymbol{W}\|_{2,1} + \lambda_2\|\boldsymbol{E}\|_1 \\ &\quad - \frac{1}{2\beta}\|\boldsymbol{Y}\|_F^2 + \frac{\beta}{2}\|\boldsymbol{X} - \mathbf{P}\boldsymbol{W}^T\mathbf{X} - \mathbf{E} + \mathbf{U}\|_F^2. \end{aligned} \quad (19)$$

where $\boldsymbol{Y}$ is Lagrange multipliers, $\boldsymbol{U} = \frac{\boldsymbol{Y}}{\beta}$ and β is a penalty parameter. The iterative algorithm updating the parameters by minimizing the Lagrangian function is shown as follows in detail.

Step 1: Update $\boldsymbol{W}$ with $\boldsymbol{P}$, $\boldsymbol{E}$ fixed by solving the problem of $\frac{\partial L}{\partial \boldsymbol{W}} = 0$:

$$\begin{aligned} \mathrm{L}(\mathbf{W}) = Tr(\boldsymbol{W}^T(S_w - \lambda S_b)\boldsymbol{W}) + \lambda_1\|\boldsymbol{W}\|_{2,1} + \lambda_2\|\boldsymbol{E}\|_1 \\ + \frac{\beta}{2}\|\boldsymbol{X} - \boldsymbol{P}\boldsymbol{W}^T\boldsymbol{X} - \boldsymbol{E} + \boldsymbol{U}\|_F^2. \end{aligned} \quad (20)$$

let $\boldsymbol{M} = \boldsymbol{X} - \boldsymbol{E} + \boldsymbol{U}$, by solving the problem of $\frac{\partial L}{\partial \boldsymbol{W}} = 0$, where

$$\frac{\partial L}{\partial \boldsymbol{W}} = 2(S_w - \lambda s_b)\boldsymbol{W} + \lambda_1\boldsymbol{D}\boldsymbol{W} + \boldsymbol{\beta}(\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W} - \boldsymbol{X}\boldsymbol{M}^T\boldsymbol{P}). \quad (21)$$

Step 2: Update $\boldsymbol{P}$ with $\boldsymbol{W}$, $\boldsymbol{E}$ fixed by solving the problem of (22),

$$\begin{aligned} &\arg\min_{\boldsymbol{P}} \frac{\beta}{2}\|\boldsymbol{X} - \mathbf{P}\boldsymbol{W}^T\boldsymbol{X} - \boldsymbol{E} + \boldsymbol{U}\|_F^2 \\ &= \arg\min_{\boldsymbol{P}} \frac{\beta}{2}\|\boldsymbol{M} - \boldsymbol{P}\boldsymbol{W}^T\boldsymbol{X}\|_F^2 \\ &\leftrightarrow \arg\min_{\boldsymbol{P}} \frac{\beta}{2} Tr(\boldsymbol{M}^T\boldsymbol{M} - 2\boldsymbol{M}^T\boldsymbol{P}\boldsymbol{W}^T\boldsymbol{X}) \\ &\leftrightarrow \arg\max_{\boldsymbol{P}} \frac{\beta}{2} Tr(\boldsymbol{M}^T\boldsymbol{P}\boldsymbol{W}^T\boldsymbol{X}) \\ &\leftrightarrow \max_{\boldsymbol{P}} Tr(\boldsymbol{P}^T\boldsymbol{M}\boldsymbol{X}^T\boldsymbol{W}). \end{aligned} \quad (22)$$

Let $SVD(\boldsymbol{M}\boldsymbol{X}^T\boldsymbol{W}) = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$, therefore $\boldsymbol{P}$ is obtained by $\boldsymbol{U}\boldsymbol{V}^T$, where *SVD* denotes the operation of singular value decomposition.

Step 3: Update $\boldsymbol{E}$ with $\boldsymbol{W}$, $\boldsymbol{P}$ fixed by solving the problem of (23):

$$\arg\min_{\boldsymbol{E}} \lambda_2\|\boldsymbol{E}\|_1 + \frac{\beta}{2}\|\boldsymbol{X} - \boldsymbol{P}\boldsymbol{W}^T\boldsymbol{X} - \boldsymbol{E} + \boldsymbol{U}\|_F^2 \quad (23)$$

Step 4: Update $\boldsymbol{Y}$ and β respectively by the following equation:

$$\boldsymbol{Y} = \boldsymbol{Y} + \beta(\boldsymbol{X} - \boldsymbol{P}\boldsymbol{W}^T\boldsymbol{X} - \boldsymbol{E}) \quad (24)$$
$$\beta = min(\beta_{max}, \rho\beta) \quad (25)$$

where ρ is constant. The iteration stops when the convergence conditions are met. The algorithm is presented in TABLE I in detail.

TABLE I
Detailed steps of the algorithm.

Algorithm. An Efficient Algorithm to Solve the Problem (18)
Input: data matrix X=$[\boldsymbol{x_1}, \boldsymbol{x_2}, \dots \dots, \boldsymbol{x_N}] \in \boldsymbol{R^{d\times n}}$,
label matrix Y=$[\boldsymbol{y_1}, \boldsymbol{y_2}, \dots \dots, \boldsymbol{y_N}] \in \boldsymbol{R^n}$
parameter $\lambda_1 = 10^{-5}$, $\lambda_2 = 10^{-6}$
The final dimension: m
Initialization: W=0;E=0;Y=0;P=arg $\min_{\boldsymbol{P^T}\mathbf{P}=\boldsymbol{I}} \frac{min}{P^T P=I} Tr(\boldsymbol{P^T}(\boldsymbol{S_w}-\boldsymbol{\lambda S_b})\boldsymbol{P})$,
$\beta = 0.1; \rho = 1.01; \beta_{max} = 10^5; \lambda = 10^{-4}$
repeat
update W by solving (21)
update P by solving (22)
update E by solving (23)
update $\boldsymbol{Y} \leftarrow \boldsymbol{Y} + \beta(\boldsymbol{X} - \boldsymbol{P}\boldsymbol{W}^T\boldsymbol{X} - \boldsymbol{E})$
$\beta \leftarrow min(\beta_{max}, \rho\beta)$
until convergence
Output: transformation matrix $\boldsymbol{W} \in \boldsymbol{R^{d\times m}}$ $(m \le d)$

IV. Data Preparation and Experimental Process

C. *Data Preparation*

ASHRAE 1043-RP provides monitoring data of the refrigeration Machine under various working conditions and different fault conditions [24]. The diagram of experimental device is shown in Fig. 3. The system is instrumented with

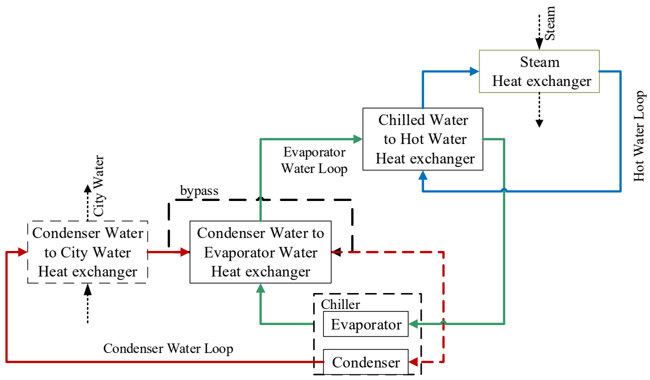


Fig. 3 The diagram of experimental device.

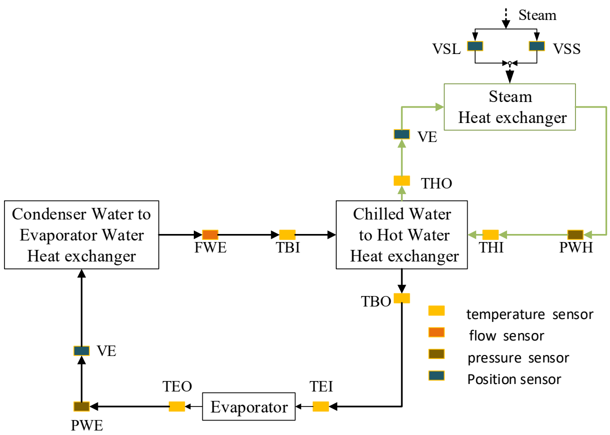


Fig. 4 Location of sensors on evaporator.

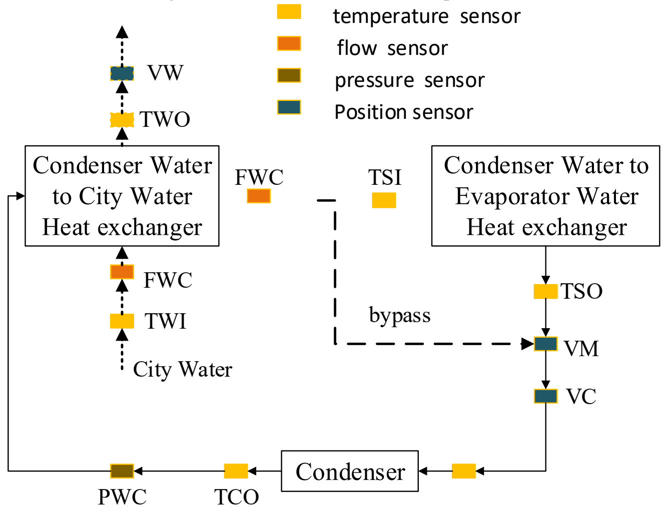


Fig. 5 Location of sensors on the condenser.

temperature, flow and position sensors in the water circuits, and pressure and temperature sensors in the refrigerant circuit [26]. Locations of the sensors in the water circuits are shown in Fig. 4 & 5.

The experiment collected a total of 64 parameters with 10s, and the data collection interval was 10s. Among them, 48 parameters were directly collected by the sensor, including 29 temperature parameters, 7 valve position parameters, 5 pressure parameters, 3 power parameters, 2 flow parameters, 2 state parameters. The remained 16 parameters were obtained through real-time calculation with VisSim software. 7 typical faults are simulated in this experiment as shown in TABLE II. In order to further analyze the characteristics of each dimension, the 64-dimensional parameters of 430 samples are visualized respectively. Fig. 6 shows the visualization of the 6-dimensional parameters of different fault conditions selected from the collected parameters.

B. Design ideal of Dimensionality Reduction

This paper designed and implemented each combination of dimensionality reductions and classification algorithm in different fault conditions, and the specific process is shown in Fig. 7.

TABLE II
Summary of faults of the refrigeration system

Health Condition	Description of Fault	Category Label
Healthy	Normal	1
False	Condenser Fouling	2
	Non-Condensables in Refrigerant	3
	Excess Oil	4
	Refrigerant Leak	5
	Refrigerant Overcharge	6
	Reduced Condenser Water Flow	7
	Reduced Evaporator Water Flow	8

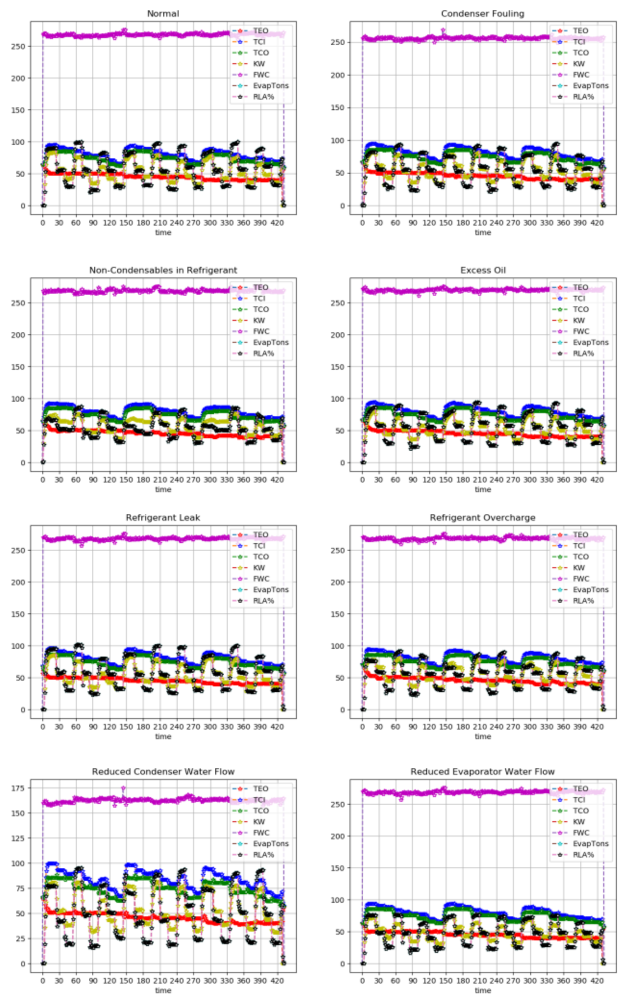


Fig. 6 Visualization of the 6-dimensional parameters in different fault conditions.

Step 1: Collect data sample of different working conditions by sensors, and build high-dimensional fault samples by feature extraction and selection.

Step 2: Normalize samples and divide them into training samples and testing samples.

Step 3: Select training samples as input of LDA, RLDA and RSLDA, and obtain transformation matrix respectively.

Step 4: The low-dimensional samples are obtained by dimensionality reduction of training and test samples through the transformation matrix.

Step 5: In order to compare the effects of dimensionality reduction processing with non-dimensionality reduction processing, take the raw samples and the obtained low-dimensional samples as the input of KNN respectively, and take accuracy of the fault diagnosis as the evaluation index of dimensionality reduction.

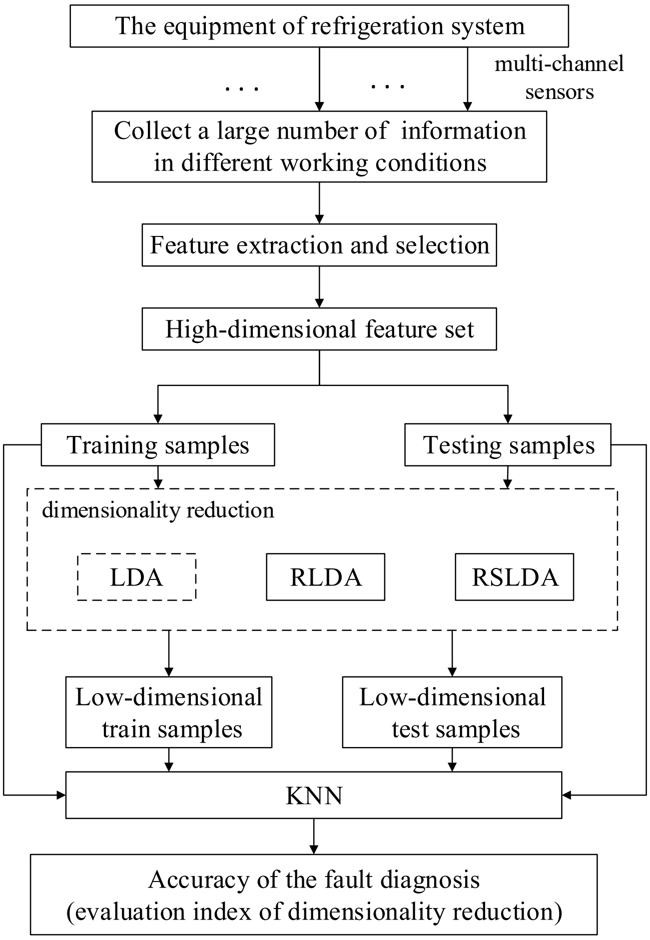


Fig. 7 Dimensionality reduction process.

V. Experimental Results and Analysis

In this section, chillers fault data set RP1043 obtained from American Society of Heating Refrigerating and Air-Conditioning Engineers are employed to evaluate the effectiveness. This paper adopted the 1-nearest-neighbor algorithm as the classifier and take its recognition rate as the index to measure the effect of dimensionality reduction. In addition, all experiments are executed on a computer with windows 10 operation system with Matlab R2016b.

A. *Discussion of Parameters*

In the experience of discussing the effect of parameters, this paper selects the first 70 percent of samples each class as the training dataset, the dimension is 7 after dimensionality reduction.

K designated by users is the initial size of analysis window in KNN. Theoretically, the smaller K is, the more precise distribution in local neighborhoods of data could be captured [23]. In order to discuss the value of parameter K, this paper selected [1,10] as the value of K. Fig. 8 shows the performance of KNN over different initial size of analysis window in dimensionality reduction methods on projected dimension 7.

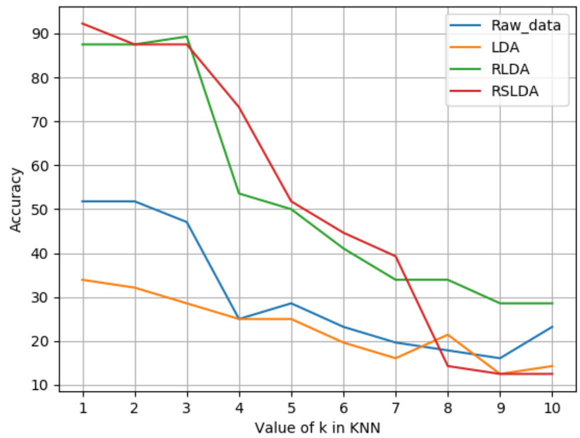


Fig. 8 The performance of KNN over different initial size of analysis window in dimensionality reduction methods on projected dimension 7.

Fig. 8 shows a clear decreasing trend with the increasing number of K. Thus, this paper take $K = 1$ as the initial size of analysis window of KNN.

λ_1 and λ_2 are two important parameters in the algorithm of RSLDA, which jointly determine the effect of dimensionality reduction. In order to discuss the values of parameters λ_1 and λ_2, this paper selected [0, 0.1] as the value range of λ_1 and λ_2. Fig. 9 shows the relationship between diagnosis accuracy and two key parameters. When values of λ_1 and λ_2 are 10^{-5} and 10^{-6}, respectively, the diagnosis accuracy reaches the higher, which is 96.43%.

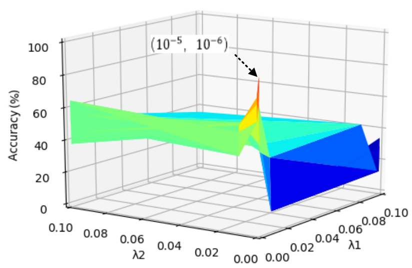
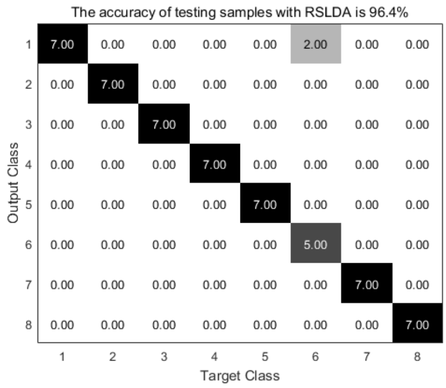


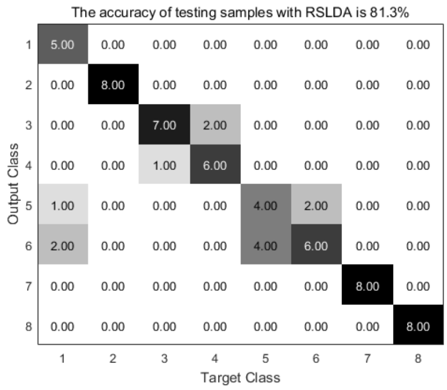
Fig. 9 Relationship between accuracy and two key parameters λ_1, λ_2.

B. Experiments on RP1043

After the parameters are determined, according to training samples with/without outliers, 2 experiments were conducted to test the performance of algorithm adopted. Fig. 10 shows the confusion matrix obtained from experiments. Experiments take the first 70 percent of samples each class as the training samples. the dimension is 7 after dimensionality reduction.



(a)



(b)

Fig. 10 Diagnosis results (confusion matrix).

Fig.10 (a) shows the confusion matrix without outliers in training samples, the diagnosis accuracy of the normal case, the condenser fouling, the non-condensables, the excess oil, the refrigerant leak, the refrigerant overcharge, the reduced condenser water flow and the reduced evaporator water flow are 62.5%, 100%, 87.5%, 75%, 50%, 75%, 100% and 100%, respectively. Fig. 10 (b) shows the confusion matrix with outliers in training samples, the diagnosis accuracy of the normal case, the condenser fouling, the non-condensables, the excess oil, the refrigerant leak, the refrigerant overcharge, the reduced condenser water flow and the reduced evaporator water flow are 100%, 100%, 100%, 100%, 100%, 71.4%, 100% and 100%, respectively.

In the following experiments on RP1043, this paper randomly selected 30, 50 and 70 percent of samples per class as the training dataset and the rest are used for the testing dataset respectively. For fair comparison with LDA, the maximum dimension is (c-1) after dimensionality reduction. Thus the final dimension is set to 2 ~(c-1) where c is the number of classes in our experiments.

(1) Experiments on RP1043 without outliers

This part compares RSLDA with LDA and RLDA on the fault dataset with 8 classification. The baseline results are acquired by using the 1-Nearest-Neighbor classifier on raw feature sets directly. This paper repeated all experiments 50 times and adopted the mean of classification accuracy as the index to measure the effect of dimensionality reduction. We recorded their classification accuracy and learning time respectively in Fig.11 and TABLE III.

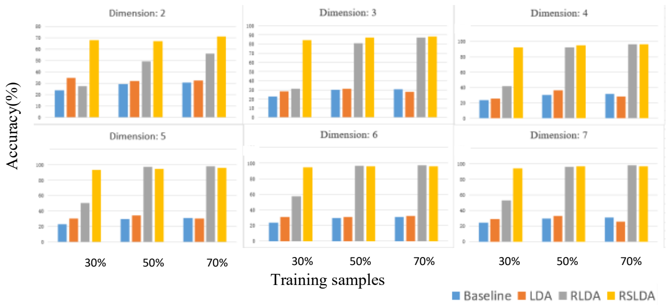


Fig. 11 The classification accuracy of the adopted method compared with other methods on 2, 3, 4, 5, 6 and 7 dimensions without outliers in training samples.

Fig. 11 exhibits that the performance of RSLDA and RLDA is almost same when training sample are 50% , 70%. But the performance of RSLDA are higher than other dimensionality reduction methods when training sample are 30%, which shows that the method is more suitable than other method for SSS problem in the paper. Table III shows that the learning time is almost same on LDA, RLDA and RSLDA.

TABLE III.
The learning time of RSLDA compared with other methods on 2, 3, 4, 5, 6 and 7 dimensions without outliers in training samples.

	Training samples	**30%**	**50%**	**70%**
Dimension	method	Time(s)	Time(s)	Time(s)
2	LDA	0.0994	0.1256	0.1034
	RLDA	0.1781	0.2347	0.2097
	RSLDA	0.1822	0.275	0.3456
3	LDA	0.1288	0.1088	0.1009
	RLDA	0.1959	0.2022	0.2009
	RSLDA	0.1172	0.2006	0.2413
4	LDA	0.1003	0.0944	0.1009
	RLDA	0.1747	0.2138	0.2053
	RSLDA	0.1066	0.0991	0.1097
5	LDA	0.0959	0.0969	0.1044
	RLDA	0.1825	0.2063	0.2197
	RSLDA	0.1081	0.1447	0.1641
6	LDA	0.0975	0.0941	0.105
	RLDA	0.18	0.2013	0.2209
	RSLDA	0.1209	0.155	0.2163
7	LDA	0.1072	0.0966	0.1094
	RLDA	0.1878	0.215	0.2106
	RSLDA	0.1341	0.1697	0.2238

(2) Experiments on RP1043 with outliers

This paper introduced 8 outliers into training sets of fault data set with 8 classification respectively and kept testing set unchanged. Similarly, all experiments are repeated 50 times and the mean of classification accuracy is adopted as the index to measure the effect of dimensionality reduction. Their classification accuracy and learning time respectively are presented in Fig. 12 and TABLE IV.

Fig. 12 exhibits that the performance of the adopted method is higher than other dimensionality reduction methods with KNN, which shows that the method is more suitable than LDA and RLDA for dimensionality reduction when training samples including outliers. TABLE IV shows the learning time is almost same on LDA, RLDA and RSLDA.

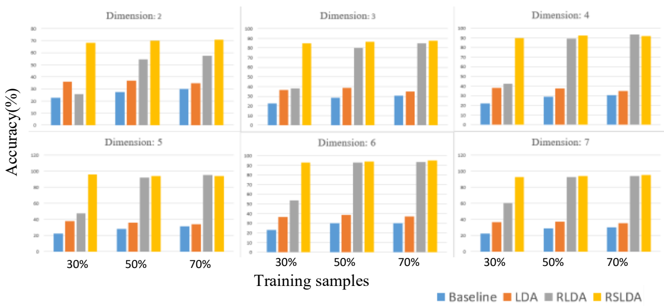


Fig. 12 The classification accuracy of the adopted method compared with other methods on 2, 3, 4, 5, 6 and 7 dimensions with outliers in training samples.

TABLE IV.
The learning time of RSLDA compared with other methods on 2, 3, 4, 5, 6 and 7 dimensions with outliers in training samples

	Training samples	**30%**	**50%**	**70%**
Dimension	method	Time(s)	Time(s)	Time(s)
2	LDA	0.0981	0.0972	0.1059
	RLDA	0.1666	0.1922	0.1931
	RSLDA	0.0978	0.0869	0.0916
3	LDA	0.0931	0.1066	0.1091
	RLDA	0.1872	0.1963	0.2138
	RSLDA	0.0938	0.1013	0.0959
4	LDA	0.1034	0.1025	0.0956
	RLDA	0.1781	0.2181	0.2069
	RSLDA	0.0944	0.1025	0.0953
5	LDA	0.0878	0.1056	0.0991
	RLDA	0.1775	0.2231	0.1888
	RSLDA	0.0959	0.1003	0.1044
6	LDA	0.0994	0.1094	0.1097
	RLDA	0.1706	0.0959	0.2306
	RSLDA	0.0969	0.1975	0.1163
7	LDA	0.1006	0.1566	0.0953
	RLDA	0.1753	0.2481	0.2081
	RSLDA	0.1016	0.12	0.1281

VI. CONCLUSIONS

In this paper, a novel efficient and robust feature selection algorithm called RSLDA, is applied to dimensionality reduction of fault data. Inspired by previous works, L_2 norm is replaced by $L_{2,1}$ norm in this algorithm to extract the most discriminant features adaptively; and an orthogonal matrix and a sparse matrix are introduced to ensure the extracted features contain the master information of the original data. Experimental results have demonstrated that the method can greatly improve the performance of fault classification when fault data set with or without outliers. In addition, this paper also proved that RSLDA is more suitable for SSS problem in fault classification.

ACKNOWLEDGMENT

This work was supported by Beijing Municipal Natural Science Foundation (No. 4192005).

REFERENCES

[1] Feng shengbo, gao linpeng, zhou fuqiu. Analysis of current situation and challenges of data center energy saving in China [J], 2015, 37(11):23-26.
[2] Li Yue-Xian, Han Zhen-Nan, Huang Hong-Chen, Ning ShaoHui. Fault diagnosis of rotating machineries based on Laplacian eigenmaps. Journal of Vibration and Shock. 2014,33(18): 21−25.
[3] Zhao Xiaoli, Zhao Rongzhen. Research on dimensionality reduction method of rotor fault data set based on global and local discriminant information fusion[J].Journal of Automation, 2017(4).
[4] K. Fukunaga, Introduction to Statistical Pattern Recognition. Cambridge, MA, USA: Academic Press, 2013.
[5] Haifeng Z , Zheng W , Feiping N . A New Formulation of Linear Discriminant Analysis for Robust Dimensionality Reduction[J]. IEEE Transactions on Knowledge and Data Engineering, 2018:1-1.
[6] I. T. Jolliffe, Principal Component Analysis. New York, NY, USA:Springer-Verlag, 1986.
[7] C. R. Rao, "The utilization of multiple measurements in problems of biological classification," J. Roy. Statistical Soc., vol. 10, no. 2,pp. 159–203, 1948.
[8] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Ann. Human Genetics, vol. 7, no. 2, pp. 179–188, 1936.
[9] Chen Sibao, Chen Daoran, Luo Bin. Maximum Spacing Criterion Based on L1-norm [J]. Journal of Electronics, 2016, 44(6):1383-1388.
[10] Sugiyama M . Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis[J]. Journal of Machine Learning Research, 2007, 8(1):1027-1061.
[11] Kwak N . Principal Component Analysis Based on L1-Norm Maximization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(9):1672-1680.
[12] Nie F , Huang H , Ding C , Luo D , Wang H . Robust Principal Component Analysis with Non-Greedy ℓ1-Norm Maximization[C]// International Joint Conference on Artificial Intelligence. 2011.
[13] Zhong F, Zhang J . Linear discriminant analysis based on L1-norm maximization [J]. IEEE Trans Image Process, 2013, 22(8):3018-3027.
[14] H Li, G Feng, J Dong, J Qiu . Fisher Discriminant Analysis with L1-Norm for Robust Palmprint Recognition[J].
[15] Lu G F,Zou J,Wang Y,Wang Z Q. Sparse L1-norm-based linear discriminant analysis [J]. Multimedia Tools and Applications, 2017.
[16] Nie F , Huang H , Cai X , Ding C. Efficient and Robust Feature Selection via Joint ℓ2, 1-Norms Minimization[C]// Advances in

Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada. Curran Associates Inc. 2010.

[17] Ren C X , Dai D Q , Yan H . Robust classification using ?2,1-norm based regression model[J]. Pattern Recognition, 2012, 45(7):2708-2718.

[18] Haifeng Z , Zheng W , Feiping N . A New Formulation of Linear Discriminant Analysis for Robust Dimensionality Reduction[J]. IEEE Transactions on Knowledge and Data Engineering, 2018:1-1.

[19] Dorfer M, Kelz R, Widmer G. Deep Linear Discriminant Analysis[J]. Nber Chapters, 2015, 5.

[20] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, “Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition,” Pattern Recognit. Lett., vol. 26, no. 2, pp. 181–191, 2005.

[21] J. Ye, “Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems,” J. Mach. Learning Res., vol. 6, no. Apr, pp. 483–502, 2005.

[22] Wen J , Fang X , Cui J , Fei L ,Yan K, et al. Robust Sparse Linear Discriminant Analysis[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2018, PP(99):1-1.

[23] Xiaoyang Li. Comparison of Feature Reduction Approaches and Classification Approachs for Pattern Recognition [R]. ResearchGate, 2016.

[24] Gao H , Huang D , Yang Y , Li L. Chinese Chunking Using ESVM-KNN[C]//International Conference on Computational Intelligence & Security. IEEE, 2007.

[25] M.C.Comstock , J.E.Braun,Development of analysis tools for the evaluation of fault detection and diagnostics in chillers,HL 99-18 Report # 4036-1,ASHRAE Research Project 1043-RP,1999

[26] Braun J E , Bendapudi S . Development and validation of a mechanistic, dynamic model for a vapor compression centrifugal liquid chiller[J]. 2002.