IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 15, 2020, accepted February 29, 2020, date of publication March 4, 2020, date of current version March 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2978287

Robust and Efficient Linear Discriminant Analysis With $L_{2,1}$-Norm for Feature Selection

LIBO YANG[1], XUEMEI LIU[1], FEIPING NIE[2], AND YANG LIU[1]

[1]School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, China
[2]Center for OPTical IMagery Analysis and Learning (OPTIMAL), School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding authors: Xuemei Liu (liuxuemei@ncwu.edu.cn) and Feiping Nie (feipingnie@gmail.com)

This work was supported in part by the National Key Research and Development Project under Grant 2017YFC0403600 and Grant 2017YFC0403604, in part by the National for Young Scientists of China under Grant 61702185, in part by the Innovation Scientists and Technicians Troop Construction Projects of Henan Province in 2014, in part by the Key Scientific and Research Project in University of Henan Province under Grant 15A520021, Grant 15A510003, and Grant 18A520034, in part by the Henan Province Science and Technology Research Program under Grant 172102210050, in part by the Open Research Foundation of Key Laboratory of Sediments in Chinese Ministry of Water Resources under Grant 2017001, and in part by the Innovation Fund for Ph.D. candidate of North China University of Water Resources and Electric Power in 2015.

ABSTRACT Feature selection and feature transformation are the two main approaches to reduce dimensionality, and they are often presented separately. In this study, a novel robust and efficient feature selection method, called FS-VLDA-L21 (feature selection based on variant of linear discriminant analysis and $L_{2,1}$-norm), is proposed by combining a new variant of linear discriminant analysis and $L_{2,1}$ sparsity regularization. Here, feature transformation and feature selection are integrated into a unified optimization objective. To obtain significant discriminative power between classes, all the data in the same class are expected to be regressed to a single vector, and the important task is to explore a transformation matrix such that the squared regression error is minimized. Therefore, we derive a new discriminant analysis from a novel view of least squares regression. In addition, we impose row sparsity on the transformation matrix through $L_{2,1}$-norm regularized term to achieve feature selection. Consequently, the most discriminative features are selected, simultaneously eliminating the redundant ones. To address the $L_{2,1}$-norm based optimization problem, we design a new efficient iterative re-weighted algorithm and prove its convergence. Extensive experimental results on four well-known datasets demonstrate the performance of our feature selection method.

INDEX TERMS Feature selection, linear discriminant analysis, $L_{2,1}$-regularization, sparsity regularization.

I. INTRODUCTION

Machine learning has been widely applied in many fields of science, such as biology, economics, sociology, and engineering. The data in these domains are always characterized by high dimensions; for example, documents, images, videos, computer vision, gene expressions, and DNA copy numbers. The time and space complexity required to process these data is extremely high. Moreover, redundant features are not only useless, but can also severely reduce the effect of machine learning [1]–[5]. Therefore, dimensionality reduction is extremely important in the data preprocessing stage [6], [7]. Several methods exist to reduce dimensionality, such as the kernel method [8]–[10], subspace projection [8], and artificial neural networks [11]. In this study, we focus on subspace projection.

Feature selection and feature transformation are two main approaches to reduce dimensionality. Feature selection is a process of selecting a subset of relevant features, and feature transformation methods transform the original features to a new feature subspace. The two methods achieve dimensionality reduction by different ways. It can be seen from the current literature that most of the literature generally focuses on one of the two approaches and few papers combine them [12]. Feature selection selects a subset of features that have significant discriminative power, and eliminates the noisy features. Thus, the selected features perform better that the original data in classification, clustering, and prediction tasks. Recently, numerous feature selection methods, such as MRMR [13], ReliefF [14], and LS [15], have been

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram.

proposed. From the perspective of search strategy, there are three models of feature selection methods: filter, wrapper and embedded models. Filter models [16]–[18] are independent of classifiers. In these models, all the features are ranked according to a predefined criterion, and the highest rankings are then selected. In the wrapper models [19]–[22], the feature subset search algorithm is wrapped around the classification model, and the usefulness of the selected features are measured based on the classifier performance. Embedded models [23]–[25] are a trade-off between the previous two models. The procedure of searching for an optimal subset of features is embedded directly in the training process. In comparison with filter models, the wrapper models and embedded models are often characterized by good performance but high computational costs. In this study, we focus on the filter models.

Linear discriminant analysis (LDA) is one of the most popular supervised dimensionality reduction methods [26], [27]. Its main objective is to obtain an optimal projection matrix, such that the ratio of the between-class distance to the within-class distance is maximized. In the past years, several variants of LDA have been proposed to achieve dimensionality reduction. S. Niijima and S. Kuhara adopted the maximum margin criterion (MMC), which is a variant of LDA, to achieve feature selection. They proposed a recursive feature selection method using the discriminant vector of the MMC [28]. Z. Zhang *et al.* proposed a Tensor Locally Linear Discriminative Analysis (TLLDA) method for image presentation [29]. Z. Zhao *et al.* elaborated a pairwise criteria based optimized LDA technique by defining new marginal inter- and intra-class scatters and proposed a variant of LDA called robust linearly optimized discriminant analysis [30]. A. Sharma *et al.* proposed a feature selection method by improving the regularized LDA technique to select important genes, crucial for the human cancer classification problem [31]. F. Yang *et al.* proposed the LDA-based feature selection method, minority class emphasized linear discriminant analysis (MCE-LDA), which addressed problems, such as singularity, overfitting, and overwhelming [32]. Zhao *et al.* proposed the soft label based LDA to achieve dimensionality reduction and applied it to image recognition and retrieval [33], [34]. Then, they integrated the Laplacian regularized least square and semi-supervised discriminant analysis into a constrained manifold regularized least square framework, and proposed a new semi-supervised dimensionality reduction method to solve the problem that underlying discriminative information cannot be fully utilized [35]. Lu *et al.* combined the structurally incoherent learning and low-rank learning with NPP to form a unified model called discriminative LR-2DNPP that could enhance the discriminative ability for feature extraction [36]. Then they proposed a robust flexible preserving embedding method. In this method, the clean data is obtained by low-rank learning and used to learn the projection matrix [37]. In this study, we propose a novel efficient and robust feature selection method based on a new variant of LDA. In this method, the objective function is defined along the idea of least squares regression, such that a transformation matrix that can minimize the loss function is obtained.

Recently, sparsity regularization has been widely investigated to achieve feature selection. A well-known regularization method is the L_1 penalty [38]. Cai *et al.* proposed the multi-cluster feature selection (MCFS), which employed the L_1-regularized regression model to select features [39]. Bradley and Mangasarian proposed the L_1-SVM method to perform feature selection using the L_1-norm regularization. The disadvantages of this method are that the number of selected features is upper bounded by the sample size and the highly correlated features are picked only one or few of them [40]. Wang *et al.* proposed a hybrid huberized support vector machine (HHSVM) method and applied it to gene selection. HHSVM combines the L_1-norm and the L_2-norm to form a more structured regularization. Thus, it performs automatic feature selection and encourages highly correlated features to be selected or eliminated together [41], [42]. Xu *et al.* proposed the $L_{1/2}$ penalty [43] and Huang *et al.* proposed the hybrid $L_{1/2+2}$ regularization (HLR) approach, which is a linear combination of the $L_{1/2}$ and L_2 penalties [44]. In this method, the $L_{1/2}$ penalty performs feature selection. Y. F. Ye *et al.* proposed robust L_{p}-norm least squares support vector regression (L_p-LSSVR) to achieve feature selection, which is robust against outliers [45]. Q. L. Ye *et al.* proposed a new discriminant method to achieve robustness by replacing the L_2-norm distances in conventional LDA with L_{p}-norm and L_{S}-norm distances [46]. Zhang *et al.* proposed the use of $L_{2,\mathrm{p}}$-norm regularization for feature selection, and presented the proximal gradient algorithm and rank-one update algorithm to solve the discrete selection problem [47]. Lu *et al.* proposed low-rank preserving projections (LRPP) for image classification. The L_{21} norm is used as a sparse constraint on the noise matrix [48]. C. Hou *et al.* proposed an unsupervised feature selection framework in which the embedding learning and sparse regression are performed simultaneously to achieve feature selection [49].

Z. Lai *et al.* proposed a series of methods based on the $L_{2,1}$-norm for linear dimensionality reduction. By replacing the L_2-norm with the $L_{2,1}$-norm to construct the objective function, these algorithms perform robust image feature extraction for classification [50]. Furthermore, they proposed a robust locally discriminant analysis via the capped norm. In this method, they constructed the robust between-class scatter matrix using the $L_{2,1}$-norm instead of the L_2-norm and imposed the $L_{2,1}$-norm regularized term on the projection matrix to ensure joint sparsity [51]. Recently, a new generalized robust regression method for jointly sparse subspace learning was proposed. This method imposes the $L_{2,1}$-norm penalty on both the loss function and the regularization term to guarantee the joint sparsity and robustness to outliers [52]. Several other studies have been conducted on the L_1-norm, $L_{1/2}$-norm, L_2-norm, L_{p}-norm ($0 < \mathrm{p} < 1$), and so on [53]–[56].

Majority of the existing sparse dimensionality reduction methods always apply an L_1-norm regularization on the transformation matrix [57], [58], enforcing sparsity on the individual elements of the transformation matrix, which does not necessarily achieve feature selection. To select lesser features, the transformation matrix needs to be forced to contain more zero rows. Therefore, in this study, we impose sparsity on the rows of the transformation matrix by adding an $L_{2,1}$ regularization term to achieve feature selection. It is difficult to optimize because the $L_{2,1}$-norm is non-smooth. Thus, an efficient iterative algorithm is proposed to solve this optimization problem. The theoretical analysis is conducted in detail and the convergence of the algorithm is proved. Extensive experiments on four real-world datasets demonstrates the effectiveness of the proposed method. The contributions of this study are summarized as follows:

1) We propose a novel efficient and robust feature selection method by combining a new variant of LDA and sparsity regularization. LDA and its variants are mostly feature transformations method. Recently sparsity regularization has been widely applied into feature selection studies. In this study, we integrate feature transformation and feature selection into a unified optimization objective to achieve feature selection.

2) We derive a new discriminant analysis for feature extraction from a novel view of least squares regression. To achieve significant discriminative power between the classes, all the data in the same class are expected to be regressed to a single vector, and the important task is to explore a transformation matrix, such that the squared regression error is minimized.

3) We impose row sparsity on the transformation matrix of the new variant of LDA through $L_{2,1}$-norm regularization to achieve feature selection. So, the most discriminative features are selected, and the redundant ones are removed simultaneously.

4) To solve the $L_{2,1}$-norm regularized optimization problem, we design an efficient iterative re-weighted algorithm. In addition, we perform the algorithm analysis and prove the convergence of the proposed algorithm.

II. LINEAR DISCRIMINANT ANALYSIS (LDA) REVIEW

LDA is a popular method of feature extraction, wherein the original high-dimensional data is transformed into low-dimensional data by the transformation matrix. The transformation process formula is

$$y = W^T x. \tag{1}$$

Here, $x \in R^d$ is the original high-dimensional data. $W \in R^{d\times m}$ is the transformation matrix ($d > m$). $y \in R^m$ is the low-dimensional data obtained after transformation.

It is well known that the main idea of LDA is that points in the same class are as close as possible, and points in different classes are as far as possible. The within-class scatter matrix S_w, between-class scatter matrix S_b, and total-class scatter matrix S_t are defined as follows:

$$S_W = \sum_{k=1}^{c} \sum_{x_i \in \pi_k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^T, \tag{2}$$

$$S_b = \sum_{k=1}^{c} n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T, \tag{3}$$

$$S_t = \sum_{x \in \pi} (x - \bar{x})(x - \bar{x})^T. \tag{4}$$

Let $X = \{x_i \in \mathbb{R}^d \,|i = 1, \cdots, n\} \in \mathbb{R}^{d\times n}$ be the given training dataset, where d is the dimensionality of the input samples and n is the number of samples. The samples are divided into c classes and each data x_i corresponds to a class label $k(1 <= k <= c)$. π_k represents the dataset of class k and n_k is the number of data points in class k. Notation $\bar{x}_k$ is the average of the data points in class k and $\bar{x}$ is the average of all the data points. $(\cdot)^T$ denotes the transpose of the matrix.

The traditional LDA solves this problem:

$$W = \arg\max_W tr((W^T S_W W)^{-1} W^T S_b W), \tag{5}$$

where $tr(\cdot)$ is the trace of the matrix. The optimal projection matrix W can be obtained by computing the eigenvectors of $S_W^{-1} S_b$ corresponding to the first m largest eigenvalues.

In our previous research work, the discriminant analysis method for feature extraction was derived by least squares regression [59]. In this feature selection method, the squared loss function is defined as

$$\varepsilon(W) = \|Y - T\|^2, \tag{6}$$

where $Y = W^T X, X = [x_1, x_2, \cdots x_n]$, $T = [t_{c1}, t_{c2}, \cdots, t_{cn}] \in \mathbb{R}^{c\times n}$, and $t_k = W^T \bar{x}_k$. $\|\cdot\|$ is the Euclidean norm, which is defined as $\|M\|^2 = tr(M^T M)$.

The goal is to minimize Eq. (6) by the linear transformation W. The associated optimization problem is

$$W = \arg\min_W \|Y - T\|^2. \tag{7}$$

To solve Eq. (7), a weighted matrix A_W is defined as

$$A_W(ij) = \begin{cases} \dfrac{1}{n_{c_i}}, & c_i = c_j, \\ 0, & otherwise. \end{cases} \tag{8}$$

A_w is an idempotent matrix. We have $A_W = A_W^2$ and $(I - A_W)^2 = I - A_W$, where I denotes the identity matrix. Thus, Eq. (2) can be rewritten as $S_W = X(I - A_W)X^T$. Then, the optimization problem in (7) can be rewritten as

$$W = \arg\min_W tr(W^T S_W W). \tag{9}$$

To avoid trivial solutions, the project matrix needs to be constrained. In this study, we focus on the condition of the orthogonal constraint. Thus, the optimization problem in (9) becomes

$$W = \arg\min_{W^T W = I} tr(W^T S_W W). \tag{10}$$

The optimal solution can be obtained by the Lagrangian function of the problem in (10). However, we propose another novel efficient and robust method to solve it.

III. FEATURE SELECTION BASED ON VARIANT OF LDA AND L2,1-NORM

In Eq. (10), S_W is the within-class scatter matrix. By substituting Eq. (2) into Eq. (10), the optimization problem becomes

$$\begin{aligned}
&\min_{W^TW=I} tr(W^T(\sum_{k=1}^{c}\sum_{x_i\in\pi_k}(x_i-\bar{x}_k)(x_i-\bar{x}_k)^T)W)\\
&=\min_{W^TW=I}\sum_{k=1}^{c}\sum_{x_i\in\pi_k} tr(W^T(x_i-\bar{x}_k)(x_i-\bar{x}_k)^TW)\\
&=\min_{W^TW=I}\sum_{k=1}^{c}\sum_{x_i\in\pi_k} tr(W^T(x_i-\bar{x}_k)(W^T(x_i-\bar{x}_k))^T)\\
&=\min_{W^TW=I}\sum_{k=1}^{c}\sum_{x_i\in\pi_k}\left\|W^T(x_i-\bar{x}_k)\right\|_2^2\\
&=\min_{W^TW=I}\frac{1}{n}\sum_{k=1}^{c}\sum_{x_i\in\pi_k}\left\|W^T(x_i-\sum_{x_i\in\pi_k}x_i)\right\|_2^2, \qquad (11)
\end{aligned}$$

where $\|\cdot\|_2$ denotes the L_2-norm defined as $\|v\|_2 = (\sum_{i=1}^{n}|v_i|^2)^{1/2}$, and vector $v\in\mathbb{R}^n$. Furthermore, the optimization problem in (11) can be rewritten as

$$\min_{W^TW=I,m_k}\sum_{k=1}^{c}\sum_{x_i\in\pi_k}\left\|W^T(x_i-m_k)\right\|_2^2. \qquad (12)$$

Note that m_k is also a variable that can be optimized. It can be easily identified that $m_k = (1/n_k)\sum_{x_i\in\pi_k}x_i$ is the optimal solution. Thus, Eq. (12) is equal to Eq. (11).

It is known that the squared loss function is extremely sensitive to outliers. To improve the robustness, we use a non-squared loss function in this study. Thus, the optimization problem in (12) becomes

$$\min_{W^TW=I,m_k}\sum_{k=1}^{c}\sum_{x_i\in\pi_k}\left\|W^T(x_i-m_k)\right\|_2, \qquad (13)$$

where Eq. (13) is not squared, and thus, the outliers have lesser importance than in Eq. (12). Then, we add the $L_{2,1}$-norm regularization term with the parameter γ to achieve feature selection. The problem becomes the following optimization problem:

$$\min_{W^TW=I,m_k}\sum_{k=1}^{c}\sum_{x_i\in\pi_k}\left\|W^T(x_i-m_k)\right\|_2+\gamma\|W\|_{2,1}. \qquad (14)$$

Solving this optimization problem is not easy because it is non-smooth. In the next section, this problem is solved using a simple and efficient algorithm.

IV. EFFICIENT ALGORITHM

A. ALGORITHM DESIGN

In this section, we propose an iterative re-weighted method to obtain solution W, such that Eq. (14) is solved. The algorithm is described in Algorithm 1, and the theoretical analysis of the algorithm is presented in the next section. In each iteration, we need to solve the following problem:

$$\min_{W^TW=I,m_k}\sum_{k=1}^{c}\sum_{x_i\in\pi_k}s_{ik}\left\|W^T(x_i-m_k)\right\|_2^2+\gamma Tr(W^TDW), \qquad (15)$$

where $s_{ik} = 1/(2\|W^T(x_i-m_k)\|)$, $d_{ii} = 1/(2\|w^i\|_2)$, and they are the weights as calculated in Algorithm 1. D is a diagonal matrix with the i-th diagonal element as d_{ii}.

Taking the derivative of Eq. (15) w.r.t. m_k and setting the derivative to zero, we obtain

$$m_k=\frac{\sum_{x_i\in\pi_k}s_{ik}x_i}{\sum_{x_i\in\pi_k}s_{ik}}. \qquad (16)$$

By substituting Eq. (16) into Eq. (15), the problem becomes

$$\begin{aligned}
&\min_{W^TW=I}\sum_{k=1}^{c}\sum_{x_i\in\pi_k}s_{ik}\left\|W^T(x_i-\frac{\sum_{x_i\in\pi_k}s_{ik}x_i}{\sum_{x_i\in\pi_k}s_{ik}})\right\|_2^2+\gamma Tr(W^TDW)\\
&=\min_{W^TW=I}\sum_{k=1}^{c}\sum_{x_i\in\pi_k}s_{ik}\left\|W^TA_{ik}\right\|_2^2\\
&\quad+\gamma Tr(W^TDW)\quad \text{s.t. } A_{ik}=(x_i-\frac{\sum_{x_i\in\pi_k}s_{ik}x_i}{\sum_{x_i\in\pi_k}s_{ik}})\\
&=\min_{W^TW=I}\sum_{k=1}^{c}\sum_{x_i\in\pi_k}s_{ik}Tr(W^TA_{ik}A_{ik}^TW)+\gamma Tr(W^TDW)\\
&=\min_{W^TW=I}Tr(W^T(\sum_{k=1}^{c}\sum_{x_i\in\pi_k}s_{ik}A_{ik}A_{ik}^T)W)+\gamma Tr(W^TDW)\\
&=\min_{W^TW=I}Tr(W^TBW)\\
&\quad+\gamma Tr(W^TDW)\quad \text{s.t. } B=\sum_{k=1}^{c}\sum_{x_i\in\pi_k}s_{ik}A_{ik}A_{ik}^T\\
&=\min_{W^TW=I}Tr(W^TMW).\quad \text{s.t. } M=(B+\gamma D) \qquad (17)
\end{aligned}$$

The columns of the optimal solution W in Eq. (17) are the l eigenvectors of M, corresponding to the first minimum l eigenvalues.

B. ALGORITHM ANALYSIS

In this section, we prove that the objective function of Eq. (14) is non-increasing in Algorithm 1. First, we consider the following lemma:

Lemma 1: For any nonzero vectors $v, v_t \in \mathbb{R}^c$, the following inequality holds:

$$\|v\|_2-\frac{\|v\|_2^2}{2\|v_t\|_2}\le\|v_t\|_2-\frac{\|v_t\|_2^2}{2\|v_t\|_2}. \qquad (18)$$

Algorithm 1 Algorithm to Solve Eq. (14)

Initialize: $s_{ik} = 1, d_{ii} = 1$.
Input: training dataset X, label information, parameters: γ, l.
Output: $W \in \mathbb{R}^{d\times l}$.
Repeat
1: Update $m_k = (\sum_{x_i\in\pi_k} s_{ik}x_i)/(\sum_{x_i\in\pi_k} s_{ik})$. Update the columns of W by the l eigenvectors of M, corresponding to the minimum l eigenvalues.
2: Update the weights $s_{ik} = 1/(2\left\|W^T(x_i - m_k)\right\|)$, $d_{ii} = 1/(2\left\|w^i\right\|_2)$.
Until converges

Proof: Obviously, the following inequality holds: $(\|v\|_2 - \|v_t\|_2)^2 \geq 0$; thus, we have

$$\begin{aligned}
&\|v\|_2^2 - 2\|v\|_2\|v_t\|_2 + \|v_t\|_2^2 \geq 0\\
&\Rightarrow 2\|v\|_2\|v_t\|_2 - \|v\|_2^2 \leq \|v_t\|_2^2\\
&\Rightarrow \|v\|_2 - \frac{\|v\|_2^2}{2\|v_t\|_2} \leq \frac{\|v_t\|_2^2}{2\|v_t\|_2}\\
&\Rightarrow \|v\|_2 - \frac{\|v\|_2^2}{2\|v_t\|_2} \leq \|v_t\|_2 - \frac{\|v_t\|_2^2}{2\|v_t\|_2}.
\end{aligned}$$

Theorem 1: Algorithm 1 monotonically decreases the objective of Eq. (14) in each iteration until the algorithm converges.

Proof: In the j-th iteration, denote the updated W and m_k by W_{J+1} and $m_{k(j+1)}$, respectively. We have

$$W_{J+1} = \underset{W^TW=I, m_k}{\arg\min} \sum_{k=1}^{c}\sum_{x_i\in\pi_k} s_{ik}\left\|W^T(x_i - m_k)\right\|_2^2 + \gamma Tr(W^TDW). \quad (19)$$

Because W_{J+1} and $m_{k(j+1)}$ are the optimal solutions of Eq. (15), the following inequality holds:

$$\begin{aligned}
&\sum_{k=1}^{c}\sum_{x_i\in\pi_k} s_{ik}\left\|W_{j+1}^T(x_i - m_{k(j+1)})\right\|_2^2 + \gamma Tr(W^TD_jW)\\
&\leq \sum_{k=1}^{c}\sum_{x_i\in\pi_k} s_{ik}\left\|W_j^T(x_i - m_{k(j)})\right\|_2^2 + \gamma Tr(W^TD_jW). \quad (20)
\end{aligned}$$

By substituting $s_{ik} = 1/(2\left\|W^T(x_i - m_k)\right\|)$, $d_{ii} = 1/(2\left\|w^i\right\|_2)$ into Eq. (20), we obtain

$$\begin{aligned}
&\sum_{k=1}^{c}\sum_{x_i\in\pi_k} \frac{\left\|W_{j+1}^T(x_i - m_{k(j+1)})\right\|_2^2}{2\left\|W^T(x_i - m_{k(j)})\right\|_2} + \gamma\sum_{i=1}^{d}\frac{\left\|w_{j+1}^i\right\|_2^2}{2\left\|w_j^i\right\|_2}\\
&\leq \sum_{k=1}^{c}\sum_{x_i\in\pi_k} \frac{\left\|W_j^T(x_i - m_{k(j)})\right\|_2^2}{2\left\|W^T(x_i - m_{k(j)})\right\|_2} + \gamma\sum_{i=1}^{d}\frac{\left\|w_j^i\right\|_2^2}{2\left\|w_j^i\right\|_2} \quad (21)
\end{aligned}$$

By substituting u and u_t in Eq. (18) with $W_{j+1}^T(x_i - m_{k(j+1)})$ and $W^T(x_i - m_{k(j)})$, respectively, we arrive at

$$\begin{aligned}
&\left\|W_{j+1}^T(x_i - m_{k(j+1)})\right\|_2 - \frac{\left\|W_{j+1}^T(x_i - m_{k(j+1)})\right\|_2^2}{2\left\|W^T(x_i - m_{k(j)})\right\|_2}\\
&\leq \left\|W_j^T(x_i - m_{k(j)})\right\|_2 - \frac{\left\|W_j^T(x_i - m_{k(j)})\right\|_2^2}{2\left\|W^T(x_i - m_{k(j)})\right\|_2}. \quad (22)
\end{aligned}$$

Thus, the following inequality holds:

$$\begin{aligned}
&\sum_{k=1}^{c}\sum_{x_i\in\pi_k}\left\|W_{j+1}^T(x_i - m_{k(j+1)})\right\|_2\\
&\quad - \sum_{k=1}^{c}\sum_{x_i\in\pi_k}\frac{\left\|W_{j+1}^T(x_i - m_{k(j+1)})\right\|_2^2}{2\left\|W^T(x_i - m_{k(j)})\right\|_2}\\
&\leq \sum_{k=1}^{c}\sum_{x_i\in\pi_k}\left\|W_j^T(x_i - m_{k(j)})\right\|_2\\
&\quad - \sum_{k=1}^{c}\sum_{x_i\in\pi_k}\frac{\left\|W_j^T(x_i - m_{k(j)})\right\|_2^2}{2\left\|W^T(x_i - m_{k(j)})\right\|_2}. \quad (23)
\end{aligned}$$

Similarly, by substituting v and v_t in Eq. (18) with w_{j+1}^i and w_j^i, respectively, we arrive at

$$\left\|w_{j+1}^i\right\|_2 - \frac{\left\|w_{j+1}^i\right\|_2^2}{2\left\|w_j^i\right\|_2} \leq \left\|w_j^i\right\|_2 - \frac{\left\|w_j^i\right\|_2^2}{2\left\|w_j^i\right\|_2}. \quad (24)$$

Parameter $\gamma > 0$; thus, the following inequality holds:

$$\begin{aligned}
&\gamma\sum_{i=1}^{d}\left\|w_{j+1}^i\right\|_2 - \gamma\sum_{i=1}^{d}\frac{\left\|w_{j+1}^i\right\|_2^2}{2\left\|w_j^i\right\|_2} \leq \gamma\sum_{i=1}^{d}\left\|w_j^i\right\|_2\\
&\quad - \gamma\sum_{i=1}^{d}\frac{\left\|w_j^i\right\|_2^2}{2\left\|w_j^i\right\|_2}. \quad (25)
\end{aligned}$$

By summing Eqs. (21), (23), and (25) on both sides, we obtain

$$\begin{aligned}
&\sum_{k=1}^{c}\sum_{x_i\in\pi_k}\left\|W_{j+1}^T(x_i - m_{k(j+1)})\right\|_2 + \gamma\sum_{i=1}^{d}\left\|w_{j+1}^i\right\|_2\\
&\leq \sum_{k=1}^{c}\sum_{x_i\in\pi_k}\left\|W_j^T(x_i - m_{k(j)})\right\|_2 + \gamma\sum_{i=1}^{d}\left\|w_j^i\right\|_2, \quad (26)
\end{aligned}$$

i.e.,

$$\begin{aligned}
&\sum_{k=1}^{c}\sum_{x_i\in\pi_k}\left\|W_{j+1}^T(x_i - m_{k(j+1)})\right\|_2 + \gamma\left\|W_{j+1}\right\|_{2,1}\\
&\leq \sum_{k=1}^{c}\sum_{x_i\in\pi_k}\left\|W_j^T(x_i - m_{k(j)})\right\|_2 + \gamma\left\|W_j\right\|_{2,1}. \quad (27)
\end{aligned}$$

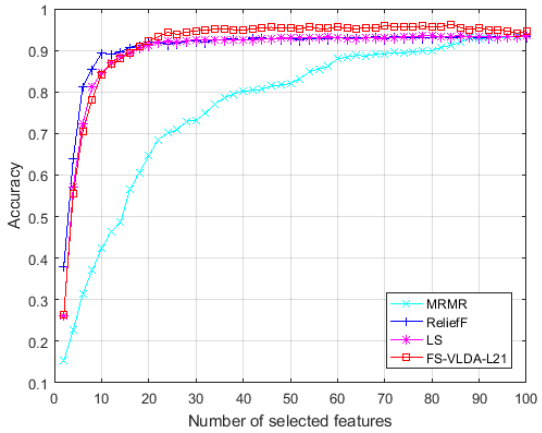

FIGURE 1. Classification accuracy comparisons between FS-VLDA-L21 and other methods on ORL data set (4 samples from per class).

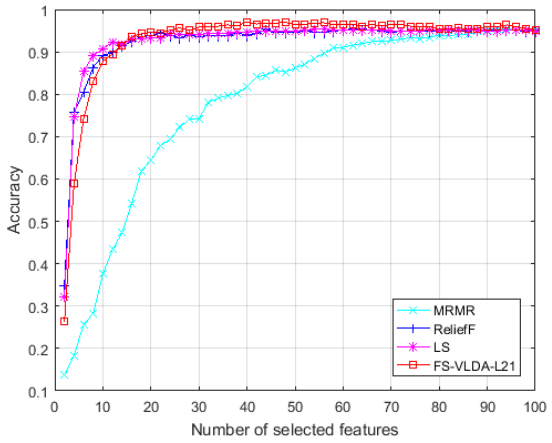

FIGURE 2. Classification accuracy comparisons between FS-VLDA-L21 and other methods on ORL data set (6 samples from per class).

Thus, Algorithm 1 monotonically decreases the objective of the problem in Eq. (14) in each iteration. Because the objective function has lower bounds, Algorithm 1 converges. Therefore, Algorithm 1 monotonically decreases the objective of the problem in Eq. (14) in each iteration until the algorithm converges.

V. EXPERIMENTAL METHOD

In this section, experiments are conducted to evaluate the performance of our proposed algorithm (denoted as FS-VLDA-L21). The comparison of our method with the previous methods, such as MRMR, ReliefF, and LS, is presented in the following section.

A. DATASET DESCRIPTION

In our experiments, four diverse public datasets, namely, ORL, YaleB, Umist, and Coil20, were used to test the performance of the different feature selection approaches.

The ORL face database included 40 distinct individuals and each individual had 10 different images. The images were captured at different times, with varied lighting, different facial expressions (open/closed eyes, smiling/not smiling), and different facial details (glasses/no glasses). The original size of each image was 112×92 pixels, with 256 grey-levels. In our experiments, we resized each image to 28×23 pixels.

The YaleB database contained a total of 2,432 face images in 38 distinct subjects. Each subject had approximately 64 near frontal images under different illuminations. The images were cropped and resized to 32×28 pixels.

The Umist database had 575 total face images of 20 different people. The original size of each image was 112×92 pixels. In our experiments, they were cropped and resized to 28×23 pixels.

The Coil20 database was composed of 1,440 images of 20 different objects. The images of each object were taken 5° apart as the object was rotated, and each object had 72 images. Each image was resized to 32×32 pixels.

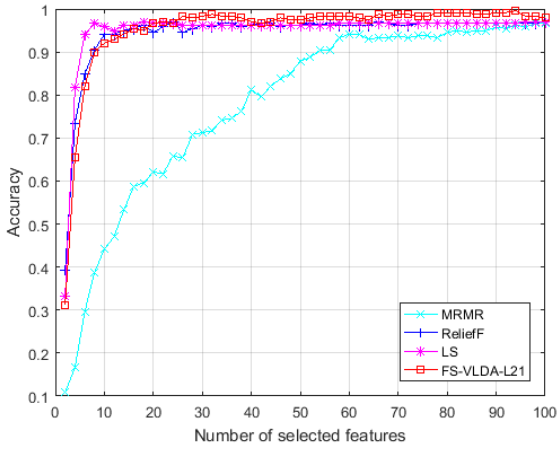

FIGURE 3. Classification accuracy comparisons between FS-VLDA-L21 and other methods on ORL data set (8 samples from per class).

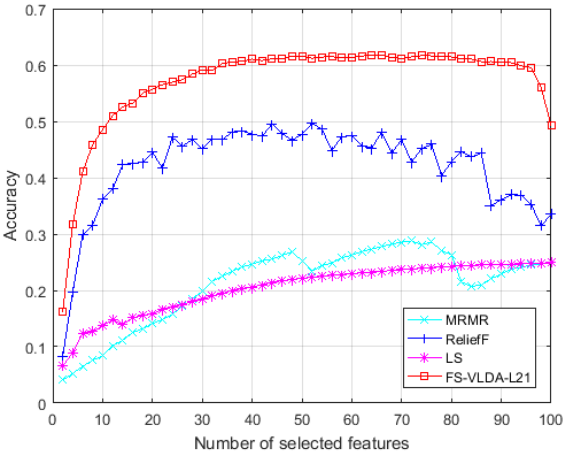

FIGURE 4. Classification accuracy comparisons between FS-VLDA-L21 and other methods on YaleB data set (4 samples from per class).

B. EXPERIMENTAL PROCESS

In each public dataset, we constructed three training datasets consisting of 4, 6, or 8 samples from each class, respectively.

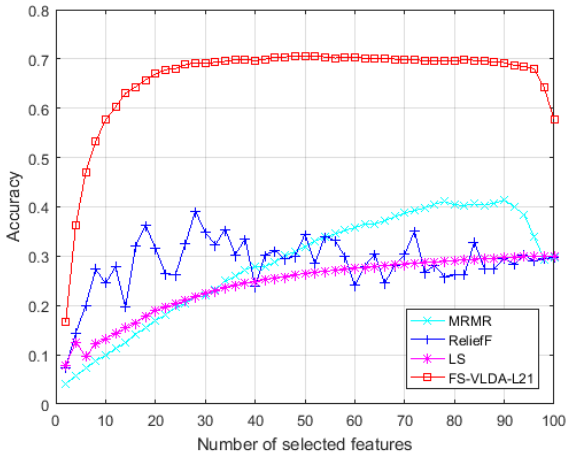

FIGURE 5. Classification accuracy comparisons between FS-VLDA-L21 and other methods on YaleB data set (6 samples from per class).

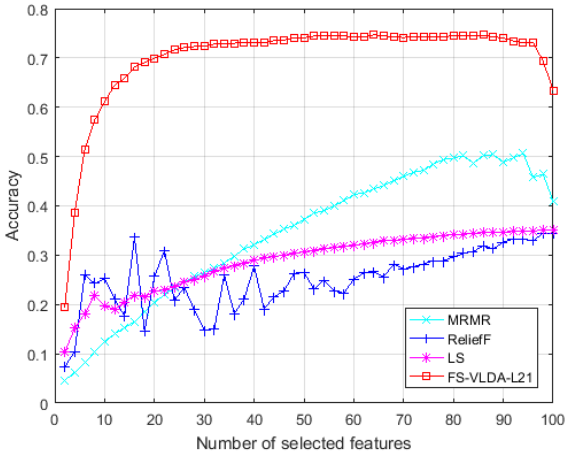

FIGURE 6. Classification accuracy comparisons between FS-VLDA-L21 and other methods on YaleB data set (8 samples from per class).

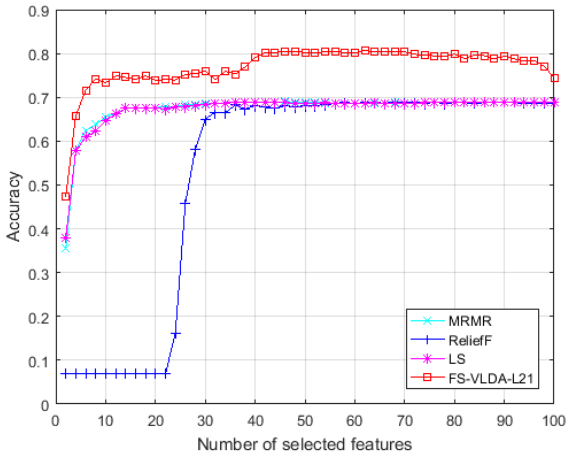

FIGURE 7. Classification accuracy comparisons between FS-VLDA-L21 and other methods on Umist data set (4 samples from per class).

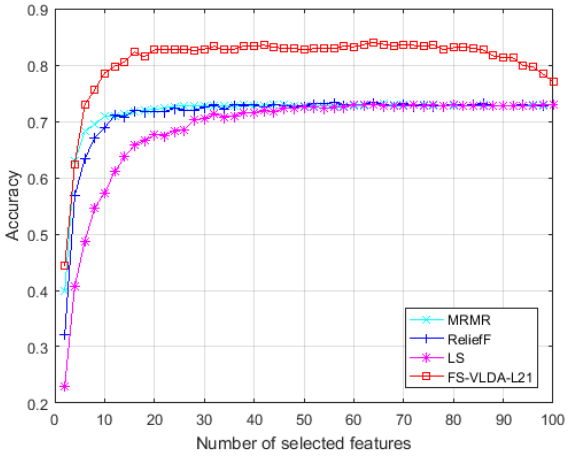

FIGURE 8. Classification accuracy comparisons between FS-VLDA-L21 and other methods on Umist data set (6 samples from per class).

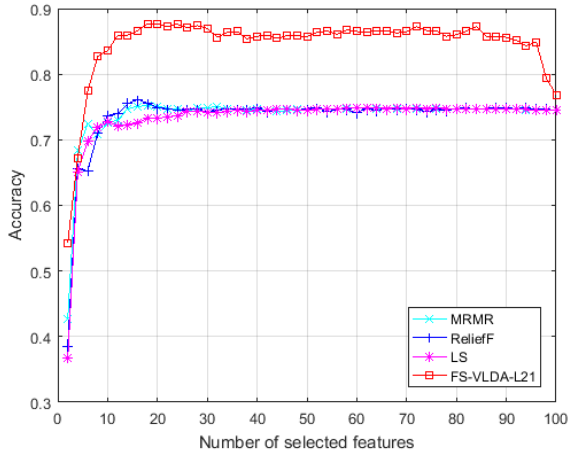

FIGURE 9. Classification accuracy comparisons between FS-VLDA-L21 and other methods on Umist data set (8 samples from per class).

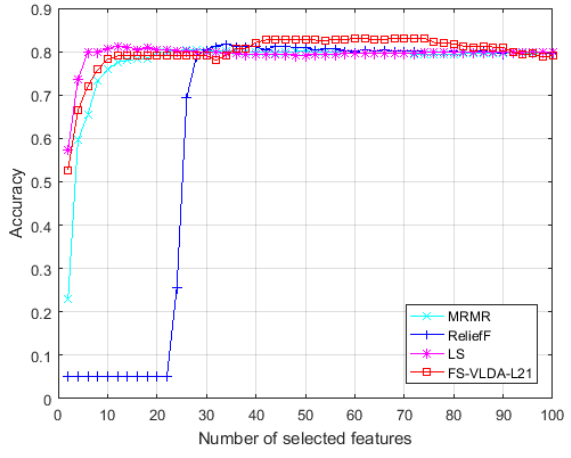

FIGURE 10. Classification accuracy comparisons between FS-VLDA-L21 and other methods on Coil20 data set (4 samples from per class).

These samples were randomly selected each time, and the remaining samples were used for testing correspondingly. The regularization parameter γ controls the tradeoff between the variant of LDA and the row sparsity of W. It plays an important role. We set the value of γ as$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}, 10^{4}, 10^{5}\}$. The value of γ

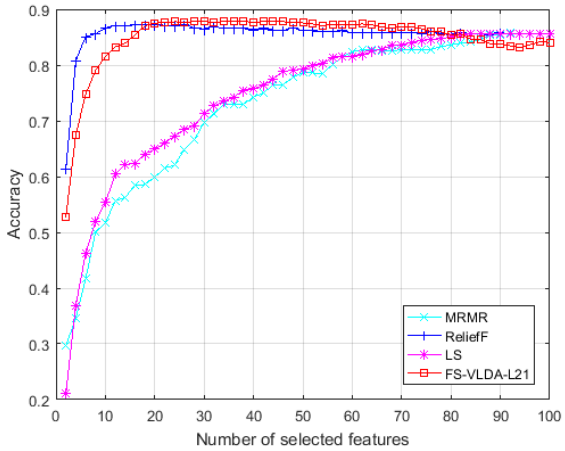

FIGURE 11. Classification accuracy comparisons between FS-VLDA-L21 and other methods on Coil20 data set (6 samples from per class).

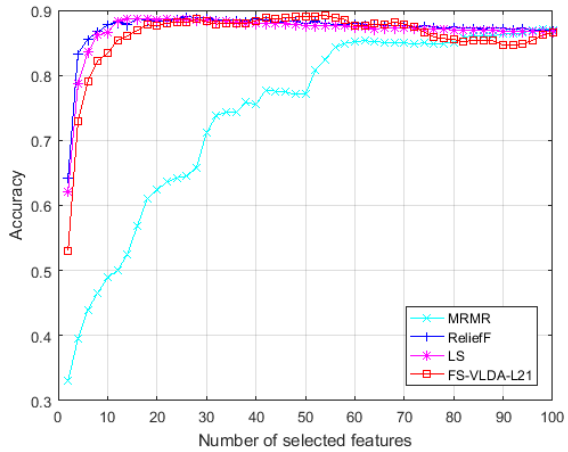

FIGURE 12. Classification accuracy comparisons between FS-VLDA-L21 and other methods on Coil20 data set (8 samples from per class).

TABLE 1. Classification Accuracy (%) of 1-nearest neighbor classifier for top 20 features (4 samples are randomly selected from each class for training).

Data set	mRMR	ReliefF	LS	FS-VLDA-L21
ORL	62.92	91.33	91.17	91.75
YaleB	15.75	39.57	16.71	55.38
Umist	69.29	6.87	69.33	79.27
Coil20	81.10	5.00	82.06	77.21
Average	57.27	35.69	64.82	75.90

that can make the experimental result optimal is adopted. To improve the efficiency of the experiment, we preprocessed the data using the PCA method. The discriminative features were selected by the traditional MRMR, ReliefF, LS, and our FS-VLDA-L21 method, respectively. The 1-nearest neighbor classifier was used to perform the classification.

TABLE 2. Classification Accuracy (%) of 1-nearest neighbor classifier for top 40 features (4 samples are randomly selected from each class for training).

Data set	mRMR	ReliefF	LS	FS-VLDA-L21
ORL	77.67	93.50	93.25	94.50
YaleB	24.55	41.10	20.44	59.63
Umist	69.66	71.23	69.70	81.37
Coil20	81.53	82.51	80.43	81.82
Average	63.35	72.09	65.95	79.33

TABLE 3. Classification Accuracy (%) of 1-nearest neighbor classifier for top 60 features (4 samples are randomly selected from each class for training).

Data set	mRMR	ReliefF	LS	FS-VLDA-L21
ORL	86.92	93.17	93.25	95.08
YaleB	32.23	40.69	22.52	60.09
Umist	70.34	69.82	70.30	81.17
Coil20	80.90	81.49	80.49	82.90
Average	67.60	71.29	66.64	79.81

TABLE 4. Classification Accuracy (%) of 1-nearest neighbor classifier for top 80 features (4 samples are randomly selected from each class for training).

Data set	mRMR	ReliefF	LS	FS-VLDA-L21
ORL	91.75	93.58	93.58	95.50
YaleB	23.05	38.33	23.89	60.50
Umist	70.38	70.10	70.38	80.12
Coil20	80.71	81.00	80.78	83.03
Average	66.47	70.75	67.16	79.79

Each experiment was performed multiple times on different random samples and the average accuracy was calculated and recorded. When the feature selection method performed better, the classification accuracy was higher.

VI. EXPERIMENTAL RESULTS

In this section, the experimental results are shown in Figs.1-12 and Tables 1-12. The Comparison and analysis of the experimental results are presented in the following section.

Figs. 1–12 depict the classification accuracies computed by the 1-nearest neighbor classifier for the four public datasets and three different training samples per dataset using different feature selection algorithms. As illustrated in these figures, for the datasets ORL and Coil20, all the methods achieve higher classification accuracy with more features selected, and more often than not, the proposed method FS-VLDA-L21 performs better than the other approaches.

TABLE 5. Classification Accuracy (%) of 1-nearest neighbor classifier for top 20 features (6 samples are randomly selected from each class for training).

Data set	mRMR	ReliefF	LS	FS-VLDA-L21
ORL	62.88	94.50	94.88	93.25
YaleB	15.55	28.25	19.07	65.96
Umist	73.41	73.89	69.85	84.26
Coil20	68.15	86.02	70.83	86.20
Average	55.00	70.66	63.66	82.42

TABLE 6. Classification Accuracy (%) of 1-nearest neighbor classifier for top 40 features (6 samples are randomly selected from each class for training).

Data set	mRMR	ReliefF	LS	FS-VLDA-L21
ORL	84.88	95.13	95.50	97.00
YaleB	26.97	30.15	24.26	69.61
Umist	74.29	73.98	73.80	83.78
Coil20	75.71	85.14	80.05	86.61
Average	65.46	71.10	68.40	84.25

TABLE 7. Classification Accuracy (%) of 1-nearest neighbor classifier for top 60 features (6 samples are randomly selected from each class for training).

Data set	mRMR	ReliefF	LS	FS-VLDA-L21
ORL	90.75	95.38	95.63	97.50
YaleB	34.43	36.57	27.05	69.78
Umist	73.93	74.42	73.76	83.47
Coil20	82.70	84.97	83.62	86.45
Average	70.45	72.83	70.01	84.30

TABLE 8. Classification Accuracy (%) of 1-nearest neighbor classifier for top 80 features (6 samples are randomly selected from each class for training).

Data set	mRMR	ReliefF	LS	FS-VLDA-L21
ORL	93.63	95.75	95.63	97.00
YaleB	37.53	30.95	28.42	69.29
Umist	74.11	74.20	74.11	83.12
Coil20	84.26	84.53	84.35	84.27
Average	72.38	71.36	70.63	83.42

TABLE 9. Classification Accuracy (%) of 1-nearest neighbor classifier for top 20 features (8 samples are randomly selected from each class for training).

Data set	mRMR	ReliefF	LS	FS-VLDA-L21
ORL	67.25	94.75	96.00	95.25
YaleB	19.66	14.24	24.78	71.16
Umist	73.30	73.35	73.40	83.86
Coil20	58.55	89.13	89.09	89.39
Average	54.69	67.87	70.82	84.91

TABLE 10. Classification Accuracy (%) of 1-nearest neighbor classifier for top 40 features (8 samples are randomly selected from each class for training).

Data set	mRMR	ReliefF	LS	FS-VLDA-L21
ORL	82.75	97.00	97.00	98.00
YaleB	31.28	18.38	28.58	73.93
Umist	74.07	73.11	74.07	84.58
Coil20	72.09	89.28	88.84	89.69
Average	65.05	69.44	72.12	86.55

TABLE 11. Classification Accuracy (%) of 1-nearest neighbor classifier for top 60 features (8 samples are randomly selected from each class for training).

Data set	mRMR	ReliefF	LS	FS-VLDA-L21
ORL	94.00	97.25	97.25	98.50
YaleB	41.06	25.15	31.84	73.90
Umist	74.07	73.98	74.12	85.20
Coil20	84.86	88.97	88.38	90.05
Average	73.50	71.34	72.90	86.91

TABLE 12. Classification Accuracy (%) of 1-nearest neighbor classifier for top 80 features (8 samples are randomly selected from each class for training).

Data set	mRMR	ReliefF	LS	FS-VLDA-L21
ORL	96.75	97.25	96.75	97.50
YaleB	49.31	30.45	33.72	73.72
Umist	74.12	74.36	74.02	85.35
Coil20	87.50	88.38	88.11	86.02
Average	76.92	72.61	73.15	85.65

For the datasets YaleB and Umist, our approach significantly outperforms the other methods.

Tables 1–12 present the detailed experimental results using the top 20, 40, 60, and 80 features for different datasets and different numbers of training samples, respectively. For each dataset, majority of the time, our method outperforms the other three methods. The last row in each table represents the average accuracy over all the datasets for each feature

selection method. On an average, the proposed method FS-VLDA-L21 performs consistently better than the other approaches in all the cases.

VII. CONCLUSION

In this study, a novel supervised feature selection method, which combines a new variant of LDA and sparsity regularization was proposed. We derived a new discriminant analysis from a novel view of least squares regression. The key work was to explore a transformation matrix such that the squared regression error was minimized. We imposed row sparsity on the transformation matrix through $L_{2,1}$-norm regularization to achieve feature selection. Therefore, feature transformation and feature selection were integrated into a unified optimization objective. Consequently, the most discriminative features were selected and the redundant ones were eliminated simultaneously. Furthermore, an efficient optimization algorithm was derived to solve the non-smooth objectives. We proved that the proposed algorithm monotonically decreased the objective until the algorithm converged. Extensive experiments were performed on four public datasets. Both theoretical analysis and empirical results demonstrated that our new feature selection method is robust, effective, and superior than the existing methods.

REFERENCES

[1] L. Wang, N. Zhou, and F. Chu, "A general wrapper approach to selection of class-dependent features," *IEEE Trans. Neural Netw.*, vol. 19, no. 7, pp. 1267–1278, Jul. 2008.

[2] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1738–1754, Nov. 2012.

[3] X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu, "Global and local structure preservation for feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 6, pp. 1083–1095, Jun. 2013.

[4] Z. Zhang, M. Zhao, and T. W. S. Chow, "Constrained large margin local projection algorithms and extensions for multimodal dimensionality reduction," *Pattern Recognit.*, vol. 45, no. 12, pp. 4466–4493, Dec. 2012.

[5] Z. Kang, H. Pan, S. C. H. Hoi, and Z. Xu, "Robust graph learning from noisy data," *IEEE Trans. Cybern.*, pp. 1–11, Dec. 2019.

[6] C. Hou, J. Wang, Y. Wu, and D. Yi, "Local linear transformation embedding," *Neurocomputing*, vol. 72, nos. 10–12, pp. 2368–2378, Jun. 2009.

[7] C. Hou, C. Zhang, Y. Wu, and F. Nie, "Multiple view semi-supervised dimensionality reduction," *Pattern Recognit.*, vol. 43, no. 3, pp. 720–730, Mar. 2010.

[8] S. Y. Kung, *Kernel Methods and Machine Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[9] T. Chanyaswad, M. Al, J. M. Chang, and S. Y. Kung, "Differential mutual information forward search for multi-kernel discriminant-component selection with an application to privacy-preserving classification," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Tokyo, Japan, Sep. 2017, pp. 1–6.

[10] S.-Y. Kung, "Discriminant component analysis for privacy protection and visualization of big data," *Multimedia Tools Appl.*, vol. 76, no. 3, pp. 3999–4034, Oct. 2017.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[12] H. Tao, C. Hou, F. Nie, Y. Jiao, and D. Yi, "Effective discriminative feature selection with nontrivial solution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 796–808, Apr. 2016.

[13] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[14] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, nos. 1–2, pp. 23–69, Oct. 2003.

[15] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2006, pp. 507–514.

[16] P. Langley, "Selection of relevant features in machine learning," in *Proc. AAAI Fall Symp. Relevance*, 1994, pp. 140–144.

[17] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford Univ. Press, 1996.

[18] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 1151–1157.

[19] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, Dec. 1997.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.

[21] A. Rakotomamonjy, "Variable selection using SVM-based criteria," *J. Mach. Learn. Res.*, vol. 3, pp. 1357–1370, Mar. 2003.

[22] C. Constantinopoulos, M. K. Titsias, and A. Likas, "Bayesian feature and model selection for Gaussian mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 1013–1018, Jun. 2006.

[23] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero-norm with linear models and kernel methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1439–1461, Mar. 2003.

[24] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, "Clustering-guided sparse structural learning for unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2138–2150, Sep. 2014.

[25] S. Wang, J. Tang, and H. Liu, "Embedded unsupervised feature selection," in *Proc. 29th AAAI Conf. Artif. Intell.*, Austin, TX, USA, 2015 pp. 470–476.

[26] K. Fukunaga, *Statistical Pattern Recognition*, 2nd ed. San Diego, CA, USA: Academic, 1990.

[27] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[28] S. Niijima and S. Kuhara, "Recursive gene selection based on maximum margin criterion: A comparison with SVM-RFE," *BMC Bioinf.*, vol. 7, no. 1, p. 543, Dec. 2006.

[29] Z. Zhang and W. S. Chow, "Tensor locally linear discriminative analysis," *IEEE Signal Process. Lett.*, vol. 18, no. 11, pp. 643–646, Nov. 2011.

[30] Z. Zhang and T. W. S. Chow, "Robust linearly optimized discriminant analysis," *Neurocomputing*, vol. 79, pp. 140–157, Mar. 2012.

[31] A. Sharma, K. K. Paliwal, S. Imoto, and S. Miyano, "A feature selection method using improved regularized linear discriminant analysis," *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 775–786, Nov. 2014.

[32] F. Yang, K. Z. Mao, G. K. K. Lee, and W. Tang, "Emphasizing minority class in LDA for feature subset selection on high-dimensional small-sized problems," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 1, pp. 88–101, Jan. 2015.

[33] M. Zhao, Z. Zhang, T. W. S. Chow, and B. Li, "A general soft label based linear discriminant analysis for semi-supervised dimensionality reduction," *Neural Netw.*, vol. 55, pp. 83–97, Jul. 2014.

[34] M. Zhao, Z. Zhang, T. W. S. Chow, and B. Li, "Soft label based linear discriminant analysis for image recognition and retrieval," *Comput. Vis. Image Understand.*, vol. 121, pp. 86–99, Apr. 2014.

[35] M. Zhao, T. W. S. Chow, Z. Wu, Z. Zhang, and B. Li, "Learning from normalized local and global discriminative information for semi-supervised regression and dimensionality reduction," *Inf. Sci.*, vol. 324, pp. 286–309, Dec. 2015.

[36] Y. Lu, Z. Lai, X. Li, W. K. Wong, C. Yuan, and D. Zhang, "Low-rank 2-D neighborhood preserving projection for enhanced robust image representation," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1859–1872, May 2018.

[37] Y. Lu, W. K. Wong, Z. Lai, and X. Li, "Robust flexible preserving embedding," *IEEE Trans. Cybern.*, to be published.

[38] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, Dec. 1996.

[39] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Washington, DC, USA, 2010, pp. 333–342.

[40] P. Bradley and O. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. 15th Int. Conf. Mach. Learn. (ICML)*, Madison, WI, USA, 1998, pp. 82–90.

[41] L. Wang, J. Zhu, and H. Zou, "Hybrid huberized support vector machines for microarray classification," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, Corvalis, OR, USA, 2007, pp. 983–990.
[42] L. Wang, J. Zhu, and H. Zou, "Hybrid huberized support vector machines for microarray classification and gene selection," *Bioinformatics*, vol. 24, no. 3, pp. 412–419, Jan. 2008.
[43] Z. Xu, H. Zhang, Y. Wang, X. Chang, and Y. Liang, "$L_{1/2}$ regularization," *Sci. China Inf. Sci.*, vol. 53, no. 6, pp. 1159–1169, Jun. 2010.
[44] H. Huang, X. Liu, and Y. Liang, "Feature selection and cancer classification via sparse logistic regression with the hybrid $L_{1/2+2}$ regularization," *PLoS ONE*, vol. 11, no. 5, May 2016, Art. no. e0149675.
[45] Y.-F. Ye, Y.-H. Shao, N.-Y. Deng, C.-N. Li, and X.-Y. Hua, "Robust lp -norm least squares support vector regression with feature selection," *Appl. Math. Comput.*, vol. 305, no. 15, pp. 32–52, Jul. 2017.
[46] Q. Ye, L. Fu, Z. Zhang, H. Zhao, and M. Naiem, "Lp-and Ls-norm distance based robust linear discriminant analysis," *Neural Netw.*, vol. 105, pp. 393–404, Sep. 2018.
[47] M. Zhang, C. Ding, Y. Zhang, and F. Nie, "Feature selection at the discrete limit," in *Proc. 28th AAAI Conf. Artif. Intell.*, Quebec City, QC, Canada, 2014.
[48] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, and C. Yuan, "Low-rank preserving projections," *IEEE Trans. Cybern.*, vol. 46, no. 8, pp. 1900–1913, Aug. 2015.
[49] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint embedding learning and sparse regression: A framework for unsupervised feature selection," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 793–804, Jun. 2014.
[50] Z. Lai, Y. Xu, J. Yang, L. Shen, and D. Zhang, "Rotational invariant dimensionality reduction algorithms," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3733–3746, Nov. 2017.
[51] Z. Lai, N. Liu, L. Shen, and H. Kong, "Robust locally discriminant analysis via capped norm," *IEEE Access*, vol. 7, pp. 4641–4652, Dec. 2019.
[52] Z. Lai, D. Mo, J. Wen, L. Shen, and W. K. Wong, "Generalized robust regression for jointly sparse subspace learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 756–772, Mar. 2019.
[53] R. Chartrand,"Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 707–710, Oct. 2007.
[54] S. Foucart and M. J. Lai, "Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0< q \ll 1$," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 395–407, May 2009.
[55] R. Chartrand,"Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data," in *Proc. IEEE Int. Symp. Biomed. Imag., From Nano Macro*, Jun. 2009, pp. 262–265.
[56] Z. Xu, X. Chang, F. Xu, and H. Zhang, "$L_{1/2}$ regularization: A thresholding representation theory and a fast solver," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1013–1027, Jul. 2012.
[57] X. Chen, J. Yang, and Z. Jin, "An improved linear discriminant analysis with L1-norm for robust feature extraction," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Stockholm, Sweden, Aug. 2014, pp. 1585–1590.
[58] F. Zhong and J. Zhang, "Linear discriminant analysis based on L1-norm maximization," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3018–3027, Aug. 2013.
[59] F. Nie, S. Xiang, Y. Liu, C. Hou, and C. Zhang, "Orthogonal vs. Uncorrelated least squares discriminant analysis for feature extraction," *Pattern Recognit. Lett.*, vol. 33, no. 5, pp. 485–491, Apr. 2012.

LIBO YANG received the B.S. and M.S. degrees in computer science from the North China University of Water Resources and Electric Power, Zhengzhou, China, in 2004 and 2011, respectively, where he is currently pursuing the Ph.D. degree. His current research interests include machine learning, pattern recognition, data mining, and information management.

XUEMEI LIU received the M.S. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 1992, and the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2008. Then, she spent several years as a Postdoctoral Researcher in the Beijing University of Aeronautics and Astronautics. She is currently a Professor and a Doctoral Supervisor of the North China University of Water Resources and Electric Power. Her current research interests include data mining, machine learning, virtual reality, and smart water.

FEIPING NIE received the Ph.D. degree in computer science from Tsinghua University. He was a Postdoctoral Research Associate, a Research Assistant Professor, and a Research Professor with The University of Texas at Arlington, Arlington, TX, USA, from 2009 to 2015. He is currently a Professor with Northwestern Polytechnical University. He has authored 160 technical articles in refereed journals and proceedings, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the *International Journal of Computer Vision*, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, ICCV, CVPR, ICML, AAAI, IJCAI, and NIPS. His current research interests include machine learning and its application fields, such as pattern recognition, data mining, computer vision, image processing, and information retrieval.

YANG LIU received the M.S. degree in computer science from the Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China, in 2012. She is currently an Associate Professor and a Master's Supervisor of the North China University of Water Resources and Electric Power. She has authored 30 technical articles in refereed journals and proceedings. Among these articles, 20 have been included by EI and SCI. Her current research interests include machine learning and its application fields, such as data mining, intelligent information process, and intelligent water conservancy.