IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 2, 2021, accepted February 20, 2021, date of publication February 24, 2021, date of current version March 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3062046

Self-Weighted Supervised Discriminative Feature Selection via Redundancy Minimization

HAIHONG YU[1,2], LIANGLIANG ZHANG[1,3], AND ZHANSHAN LI[1,2]

[1]Key Laboratory for Symbol Computation and Knowledge Engineering of National Education Ministry, Jilin University, Changchun 130012, China
[2]College of Computer Science and Technology, Jilin University, Changchun 130012, China
[3]College of Software, Jilin University, Changchun 130012, China

Corresponding author: Zhanshan Li (lizs@jlu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61802056, in part by the Natural Science Foundation of Jilin Province under Grant 20180101043JC, in part by the Development and Reform Committee Foundation of Jilin province of China under Grant 2019C053-9, and in part by the Open Research Fund of Key Laboratory of Space Utilization, Chinese Academy of Sciences under Grant LSU-KFJJ-2019-08.

ABSTRACT Feature selection plays a key role in many machine learning problems. Especially as an important data preprocessing method, robust and pragmatic feature selection methods can be applied to extract meaningful features and eliminate redundant ones. As we all known, many feature selection methods select features by using some certain feature evaluation criteria to obtain the corresponding score for every feature, such that we can select high score features. Unfortunately, correlated features usually connect with each other, which may result in large correlations between top ranked features, such that the redundancy among the selected features is brought about. To solve this problem, we introduce the redundancy matrix **A** in the AGRM (a novel auto-weighted feature selection framework via global redundancy minimization) framework. Meanwhile, we introduce the adaptive redundancy matrix **S** and treat the redundancy matrix **S** as an optimizing variable, rather than setting the redundant matrix **S** as a prior. In addition, we propose a robust algorithm to efficiently address the constrained optimization problem. Finally, extensive experiments on six datasets show the superiority of our proposed method.

INDEX TERMS Feature selection, linear discriminant analysis (LDA), redundancy minimization, sparse regularization.

I. INTRODUCTION

In the age of big data, high-dimensional data have already existed widely in many fields, such as machine learning, bioinformatics and so on. However, it may result in large requirement for time and space. At the same time, many machine learning and data mining tasks may become difficult in high-dimensional data. This is the so-called the curse of dimensionality. Besides, data of high dimensionality can also make learning model overfitting, which may result in bad performance. To solve this problem, feature selection techniques are devised to select the most relevant and representative subset of features, which can greatly improve model's performance. When getting a relevant subset of features, we can use traditional data techniques for effective processing.

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

According to the availability of label information, feature selection methods can be classified into unsupervised [1]–[3], semi-supervised [4], [5] and supervised [6], [7]. From another perspective, feature selection algorithms can also be divided into three categories roughly: filter [8], [9], wrapper [10], [11], and embedded methods [12]–[14]. The difference in these categories depends on how the learning algorithm is incorporated in evaluating and selecting features. Filter methods are not linked to any learning algorithm, and select features by using the certain characteristics of data. Therefore, filter methods are independent of the specific learning algorithm, and its characteristic is the use of global statistical information. Wrapper methods are closely related to the performance of a predefined learning algorithm, which can usually get better performance under a specific feature subset. However, the computational cost is very high. Embedded methods have the advantages of filter and wrapper methods, which combine the feature selection with the model training. Moreover, embedded methods can obtain

better model's performance and reduce the requirement of time and space. Least square regression is usually applied to embedded model. Wu *et al.* [15] proposed a novel supervised feature selection method, which employed orthogonal least square regression model with feature weighting. Besides, with the development of neural network, some feature selection methods [16] have used neural network to minimize the reconstruct errors.

Recently, sparse learning theory has an important influence on the improvement of algorithms. For instance, Liu and Tsang [17], [18] applied sparse learning theory to their framework, such that the decision tree can become easier to interpret and understand in high dimensions. In addition, Liu *et al.* [19] accelerated k-means clustering by using the sparse learning technology. According to the structure of the constraints, sparsity can be got from two types of regularizers:

1) Flat sparsity. The sparsity is realized by l_1-norm or l_0-norm. Typical methods include LARS [20], linear gradient search [21];

2) Structural sparsity. Group features are selected by the $l_{2,1}$-norm or $l_{2,0}$-norm.

Many researchers have tried to apply the sparse regularization to the embedded feature selection models as well. Tibshirani [22] imposed l_1-norm regularization term on the feature selection model (Lasso). However, Lasso only works for binary classification. For multi-class feature selection problem, we prefer to structured sparsity regularization, which can select the features across all the classes with joint sparsity. Besides, due to the virtue of $l_{2,0}$-norm for its non-convex and non-smooth properties, problem optimization may become difficult. Therefore, people usually tend to consider the convex $l_{2,1}$-norm as the regularization term. Based on this, Nie *et al.* [6] proposed an effective feature selection method to select the representative features, which can be realized by imposing joint $l_{2,1}$-norm minimization on loss function and regularization term. Besides, Yang *et al.* [23] combined the discriminant analysis with $l_{2,1}$-norm minimization to address unsupervised feature selection problem. Yan *et al.* [24] performed special processing on feature weight matrix, where $l_{2,1}$-norm was introduced and the non-negative constraint was used to gain the row-sparsity. Moreover, because of the unclear meaning of the regularization parameter for the $l_{2,1}$-norm constrained problem, some work is needed to seek good regularization parameter. However, there is a better solution to avoid this additional work. We can focus on the original $l_{2,0}$-norm constrained problem, because its regularization parameter has a clear meaning, which is the number of features selected. Therefore, it is very important to find a solution to address the original $l_{2,0}$-norm constrained problem. A pragmatic and robust algorithm was proposed by Cai *et al.* [25], which was based on augmented Lagrangian method to solve the original $l_{2,0}$-norm constrained optimization problem. Because its performance is sensitive to its initialization algorithm, Pang *et al.* [26] proposed a new framework to solve the original $l_{2,0}$-norm constrained feature selection problem, which can associate the $l_{2,0}$-norm constrained feature selection problem with linear discriminant analysis (LDA). Besides, Wang *et al.* [27] proposed a novel discriminative feature selection approach by enforcing orthogonal $l_{2,0}$-norm constraint, which applied a Structured Sparse Subspace Learning module to solve subspace sparsity problems.

Due to the good discrimination ability of LDA in feature selection, Zhang *et al.* [14] made the most of the good virtue of LDA to construct optimization problem and proposed an effective and robust feature selection method by introducing the $l_{2,0}$-norm regularization term. As we all known, the problem of feature selection needs to be considered from two aspects: the discriminant ability of features and the correlation between features. In view of the feature redundancy between the selected features, Nie *et al.* [28] presented an auto-weighted feature selection framework via global redundancy minimization (AGRM). However, Wu *et al.* [29] didn't treat redundancy matrix as a prior, but set it as a variable, which can reduce the redundancy between relevant features by adaptively evaluating the correlation between features.

Regarding the main contributions, we summarized as follows:

1) we apply self-weighted linear discriminant analysis (SLDA) method and introduce $l_{2,0}$-norm to SLDA problem.

2) In order to reduce redundancy between selected features, we introduce the redundancy matrix $\mathbf{A}$ in the AGRM framework to our constrained optimization problem. Meanwhile, we introduce the adaptive redundancy matrix $\mathbf{S}$ and treat the redundancy matrix $\mathbf{S}$ as a variable, rather than setting the redundancy matrix as a prior.

3) For the optimization of the algorithm, we propose an efficient algorithm based on the augmented Lagrangian method to solve the constrained optimization problem, such that the global optimal solution can be obtained.

4) In multi-classification task, extensive experiments on six datasets show the superiority of our proposed method compared with other seven state-of-the-art feature selection methods by using two classifiers as K-NN and linear SVM.

II. SPARSE LEARNING BASED FEATURE SELECTION

For several binary feature selection methods based on sparse learning, it can be described as the following problem [25]:

$$\begin{aligned} &\min_{\mathbf{W},b} \left\| \mathbf{X}^T\mathbf{w} + b\mathbf{1} - \mathbf{y} \right\|_2^2 \\ &s.t.\ \ \|\mathbf{w}\|_0 = k \end{aligned} \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{d\times n}$ is the training data. $\mathbf{y} \in \mathbb{B}^{n\times 1}$ is the binary label. $\mathbf{w} \in \mathbb{R}^{d\times 1}$ is the learning model. $\mathbf{1} \in \mathbb{R}^{n\times 1}$ denotes a column vector whose entries are 1. k denotes the number of the selected feature.

However, optimization of the problem (1) is NP-hard. In some cases, the constraint can be relaxed with the following formulation:

$$\min_{\mathbf{W},b} \left\| \mathbf{X}^T\mathbf{w} + b\mathbf{1} - \mathbf{y} \right\|_2^2 + \lambda \|\mathbf{w}\|_0 \tag{2}$$

where $\lambda \in \mathbb{R}^+$ is the regularization parameter. However, the solution to address (2) remains challenging. To simplify the issue, we usually choose the l_1 norm instead of the l_0 norm. The problem can be described in the following form:

$$\min_{\mathbf{W},b} \left\| \mathbf{X}^T\mathbf{w} + b\mathbf{1} - \mathbf{y} \right\|_2^2 + \lambda \left\| \mathbf{w} \right\|_1 \tag{3}$$

This type of problem has been extensively studied. Furthermore, a closed form solution can be obtained through theoretical proof.

Of course, we can take some strategies to extend the above model to deal with multi-class tasks, such as one-versus-all and one-versus-one, but some structural sparsity is preferred which can select features across all the classes.

To solve the problem, Nie *et al.* [6] proposed a structured sparse regression model with imposing joint $l_{2,1}$-norm minimization on loss function and regularization term to solve the multi-class tasks. The specific description is as follows:

$$\min_{\mathbf{W}} \left\| \mathbf{X}^T\mathbf{W} - \mathbf{Y} \right\|_{2,1} + \lambda \left\| \mathbf{W} \right\|_{2,1} \tag{4}$$

where $\mathbf{Y} \in \mathbb{B}^{n\times c}$ denotes the binary label matrix and $\mathbf{W} \in \mathbb{R}^{d\times c}$ denotes the learned model.

In fact, from the sparsity perspective, we prefer the $l_{2,0}$-norm regularization term. An effective and practical feature selection approach was proposed by Cai *et al.* [25] to deal with the original $l_{2,0}$-norm constrained problem. The model is described as follows:

$$\begin{aligned} &\min_{\mathbf{W},b} \left\| \mathbf{X}^T\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y} \right\|_{2,1} \\ &s.t.\ \left\| \mathbf{W} \right\|_{2,0} = k \end{aligned} \tag{5}$$

In [25], a robust and pragmatic algorithm was used to solve the problem (5), which was based on the augmented Lagrangian method. By using their method, we can avoid spending some time to adjust the regularization parameter. Because the regularization parameter in (5) has a clear meaning which refers to the number of selected features.

III. THE PROPOSED METHOD

Based on the relevant technology in the previous section, the proposed method is presented as follows:

A. SELF-WEIGHT LINEAR DISCRIMINANT ANALYSIS (SLDA)

For training data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, data point $\mathbf{x}_i$ $(1 \le i \le n)$ has a corresponding label $\mathbf{y}_i \in \{0, 1\}^{c\times 1}$, whose corresponding label matrix is $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]^T \in \mathbb{B}^{n\times c}$. Moreover, d denotes dimension of feature and n denotes data number. Besides, the total-class scatter matrix $\mathbf{S}_t$, the between-class scatter matrix $\mathbf{S}_b$ and the within-class scatter matrix $\mathbf{S}_w$ can be expressed as:

$$\begin{cases} \mathbf{S}_t = \mathbf{X}\mathbf{H}\mathbf{X}^T \\ \mathbf{S}_b = \mathbf{X}\mathbf{H}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\mathbf{H}\mathbf{X}^T \\ \mathbf{S}_w = \mathbf{S}_t - \mathbf{S}_b = \mathbf{X}\mathbf{H}\left(\mathbf{I} - \mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\right)\mathbf{H}\mathbf{X}^T \end{cases} \tag{6}$$

where matrix $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is idempotent which has following properties:

$$\mathbf{H} = \mathbf{H}^T = \mathbf{H}\mathbf{H}^T = \mathbf{H}^T\mathbf{H} \tag{7}$$

Based on the Eq. (6), we can formulate linear discriminant analysis (LDA) as follows:

$$\begin{cases} \max\limits_{\mathbf{W}} \mathbf{Tr}\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right) = \max\limits_{\mathbf{W}} \mathbf{Tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}_b\mathbf{X}^T\mathbf{W}\right) \\ \min\limits_{\mathbf{W}} \mathbf{Tr}\left(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right) = \min\limits_{\mathbf{W}} \mathbf{Tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}_w\mathbf{X}^T\mathbf{W}\right) \end{cases} \tag{8}$$

where $\mathbf{L}_b = \mathbf{H}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\mathbf{H}$ and $\mathbf{L}_w = \mathbf{H} - \mathbf{H}\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\mathbf{H}$.

According to the problem in (8), we can get a novel self-weighted linear discriminant analysis where the optimal weight can be obtained. The description is as follows:

$$\begin{aligned} &\max_{\mathbf{W},\lambda} \lambda\, \mathbf{Tr}\left(\mathbf{W}^T\left(\mathbf{S}_b - \lambda\mathbf{S}_w\right)\mathbf{W}\right) \\ &= \max_{\mathbf{W},\lambda} \mathbf{Tr}\left(\mathbf{W}^T\mathbf{X}\left(\lambda\mathbf{L}_b - \lambda^2\mathbf{L}_w\right)\mathbf{X}^T\mathbf{W}\right) \\ &\Rightarrow \min_{\mathbf{W},\lambda} \mathbf{Tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}^{(\lambda)}\mathbf{X}^T\mathbf{W}\right) \end{aligned} \tag{9}$$

where $\mathbf{L}^{(\lambda)} = \lambda^2\mathbf{L}_w - \lambda\mathbf{L}_b$ is associated with the weight λ which is introduced as a variable be optimized in (9).

B. CONSTRUCTION OF THE REDUNDANCY MATRIX A

For training data matrix $\mathbf{X}$, $\mathbf{X}_{(i)}$ and $\mathbf{X}_{(j)}$ $(i, j = 1, 2, \ldots, d)$ denote the i-th feature, the j-th feature respectively. Thus, the i-th and the j-th centralized features can be expressed as:

$$\begin{cases} \mathbf{f}_i = \mathbf{H}_n\mathbf{X}_{(i)}^T \\ \mathbf{f}_j = \mathbf{H}_n\mathbf{X}_{(j)}^T \end{cases} \tag{10}$$

where $\mathbf{f}_i \in \mathbb{R}^{n\times 1}$ and $\mathbf{f}_j \in \mathbb{R}^{n\times 1}$ are all column vectors. Therefore, we can obtain matrix $\mathbf{A}$ as

$$\mathbf{A}_{i,j} = \left(\mathbf{B}_{i,j}\right)^2 = \left(\frac{\mathbf{f}_i^T\mathbf{f}_j}{\left\|\mathbf{f}_i\right\|\left\|\mathbf{f}_j\right\|}\right)^2 \tag{11}$$

We can see, $\mathbf{A} = \mathbf{B} \circ \mathbf{B}$, where $\circ$ is the Hadamard product. Moreover, we can further obtain the following form as

$$\mathbf{B} = \mathbf{D}\mathbf{F}^T\mathbf{F}\mathbf{D} = (\mathbf{F}\mathbf{D})^T\mathbf{F}\mathbf{D} \tag{12}$$

where $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_d]$ and $\mathbf{D}$ is a diagonal matrix, whose element is $\frac{1}{\|\mathbf{f}_i\|}$, $i = 1, 2, \ldots, d$. As can be seen from Eq. (12), the matrix $\mathbf{B}$ is positive semi-positive.

C. SELF-WEIGHTED DISCRIMINATIVE FEATURE SELECTION VIA REDUNDANCY MINIMIZATION

Although redundancy matrix $\mathbf{A}$ can measure the redundancy between features, it is not appropriate to set redundant matrix as a priori. As a result, we introduce the adaptive redundant matrix $\mathbf{S}$ and treat $\mathbf{S}$ as an optimizing variable. So, the redundancy among features can be reduced by minimizing the term $\mathbf{Tr}\left(\mathbf{W}^T(\mathbf{A} + \mathbf{S})\mathbf{W}\right)$. Besides, we make projection matrix $\mathbf{W}$ become a row-sparse matrix to gain the non-redundant features. Therefore, optimization problem of our paper can be

expressed as follows:

$$\min_{\mathbf{W},\lambda,\mathbf{S}} \frac{1}{2}\mathbf{Tr}\left(\mathbf{W}^T\left(\mathbf{X}\mathbf{L}^{(\lambda)}\mathbf{X}^T+\mathbf{A}\right)\mathbf{W}\right)+\frac{1}{2}\mathbf{Tr}\left(\mathbf{W}^T\mathbf{S}\mathbf{W}\right)$$
$$s.t.\ \|\mathbf{W}\|_{2,0}=k, \mathbf{S}\geq\mathbf{0}, \mathbf{Tr}\left(\mathbf{S}^{-1}\right)\leq 1 \tag{13}$$

where $\mathbf{W}\in\mathbb{R}^{d\times m}$ has only k nonzero rows and $\mathbf{S}\geq\mathbf{0}$ represents $\mathbf{S}$ is a positive semi-definite matrix. At the same time, constraint $\mathbf{Tr}\left(\mathbf{S}^{-1}\right)\leq 1$ is to avoid the potential trivial solution that's the zero solution of $\mathbf{S}$.

Theorem 1: For the variables $\mathbf{W}$, λ and $\mathbf{S}$, Problem (13) is convex.

Proof: In problem (13), we can see that the first term in the minimization problem is convex for all variables. For the second term in minimization problem, it can be reformulated as

$$\mathbf{Tr}\left(\mathbf{W}^T\mathbf{S}\mathbf{W}\right)=\sum_{i=1}^{m}\mathbf{w}_i^T\mathbf{S}\mathbf{w}_i$$

where $\mathbf{w}_i^T\mathbf{S}\mathbf{w}_i$ is the matrix fractional function. If $\mathbf{S}\geq\mathbf{0}$, it can be proved by [30] that the second term in the minimization problem is a convex function w.r.t. $\mathbf{w}_i$. In addition, it is proved by [30] that summation can preserve convexity, such that $\mathbf{Tr}\left(\mathbf{W}^T\mathbf{S}\mathbf{W}\right)=\sum_{i=1}^{m}\mathbf{w}_i^T\mathbf{S}\mathbf{w}_i$ is convex with respect to $\mathbf{W}$, λ, $\mathbf{S}$. Thus, problem (13) is convex for the variable $\mathbf{W}$, λ and $\mathbf{S}$. ■

IV. OPTIMIZATION ALGORITHM

A. GENERAL AUGMENTED LAGRANGIAN MULTIPLIER (ALM) METHOD

The general augmented Lagrange multipliers (ALM) method is applied to solve the following constrained optimization problems:

$$\min f(\mathbf{X}) \qquad s.t.\ \mathbf{Tr}(\varphi(\mathbf{X}))=0 \tag{14}$$

The augmented Lagrangian function of problem (14) can be described as follows:

$$L(\mathbf{X},\Lambda,\mu)=f(\mathbf{X})+\mathbf{Tr}\left(\Lambda^T\varphi(\mathbf{X})\right)+\frac{\mu}{2}\|\varphi(\mathbf{X})\|_F^2 \tag{15}$$

We can further have as follows:

$$L(\mathbf{X},\Lambda,\mu)=f(\mathbf{X})+\frac{\mu}{2}\left\|\varphi(\mathbf{X})+\frac{\Lambda}{\mu}\right\|_F^2 \tag{16}$$

where μ is a positive scalar called the quadratic penalty parameter and Λ is the Lagrangian multiplier. Accordingly, Algorithm 1 summarizes the procedure of the ALM.

B. REFORMULATION AS A CONSTRAINED PROBLEM

By virtue of ALM method, we introduce the auxiliary variable $\mathbf{V}$. Eq. (13) can be reformulated as

$$\min_{\mathbf{W},\lambda,\mathbf{V},\mathbf{S}} \mathbf{Tr}\left(\mathbf{W}^T\left(\mathbf{X}\mathbf{L}^{(\lambda)}\mathbf{X}^T+\mathbf{A}+\mathbf{S}\right)\mathbf{W}\right)+\mu\left\|\mathbf{W}-\mathbf{V}+\frac{\Lambda}{\mu}\right\|_F^2$$
$$s.t.\ \|\mathbf{V}\|_{2,0}=k, \mathbf{S}\geq\mathbf{0}, \mathbf{Tr}\left(\mathbf{S}^{-1}\right)\leq 1 \tag{17}$$

Algorithm 1 Augmented Lagrange Multiplier (ALM)

Initialization:

1: Initialize Λ.
2: Initialize $\mu>0$.
3: Initialize $\rho\geq 1$.

repeat

1: Update $\mathbf{X}$ by $\arg\min_{\mathbf{X}} f(\mathbf{X})+\frac{\mu}{2}\left\|\varphi(\mathbf{X})+\frac{\Lambda}{\mu}\right\|_F^2$
2: Update Λ by $\Lambda+\mu\varphi(\mathbf{X})$
3: Update μ by $\rho\mu$

until Converges

C. SOLVE THE ABOVE CONSTRAINED PROBLEM

To deal with the problem (17), we utilize a robust and efficient algorithm, which is based on the general ALM.

1) **Update** λ with $\mathbf{W}$, $\mathbf{V}$, $\mathbf{S}$ fixed

When taking the derivative of problem (17) w.r.t. λ and setting it to zero, we can obtain

$$\lambda=\frac{\mathbf{Tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}_b\mathbf{X}^T\mathbf{W}\right)}{2\,\mathbf{Tr}\left(\mathbf{W}^T\mathbf{X}\mathbf{L}_w\mathbf{X}^T\mathbf{W}\right)} \tag{18}$$

2) **Update** $\mathbf{S}$ with $\mathbf{W}$, $\mathbf{V}$, λ fixed

The subproblem becomes,

$$\min_{\mathbf{S}} \mathbf{Tr}\left(\mathbf{W}^T\mathbf{S}\mathbf{W}\right) \qquad s.t.\ \mathbf{S}\geq\mathbf{0}, \mathbf{Tr}\left(\mathbf{S}^{-1}\right)\leq 1 \tag{19}$$

By using Cauchy–Schwartz inequality and $\mathbf{Tr}\left(\mathbf{S}^{-1}\right)\leq 1$, we can have

$$\begin{aligned}&\mathbf{Tr}\left(\mathbf{W}^T\mathbf{S}\mathbf{W}\right)\\&=\mathbf{Tr}\left(\mathbf{S}\mathbf{W}\mathbf{W}^T\right)\times 1\\&\geq\mathbf{Tr}\left(\mathbf{S}^{\frac{1}{2}}\left(\mathbf{W}\mathbf{W}^T\right)^{\frac{1}{2}}\left(\mathbf{W}\mathbf{W}^T\right)^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}}\right)\mathbf{Tr}\left(\mathbf{S}^{-\frac{1}{2}}\mathbf{S}^{-\frac{1}{2}}\right)\\&\geq\left(\mathbf{Tr}\left(\left(\mathbf{W}\mathbf{W}^T\right)^{\frac{1}{2}}\right)\right)^2\end{aligned} \tag{20}$$

when equality holds, we have

$$\gamma\left(\mathbf{W}\mathbf{W}^T\right)^{\frac{1}{2}}\mathbf{S}^{\frac{1}{2}}=\mathbf{S}^{-\frac{1}{2}}\Rightarrow\gamma\left(\mathbf{W}\mathbf{W}^T\right)^{\frac{1}{2}}=\mathbf{S}^{-1}$$

where γ is an arbitrary constant. Besides, we can obtain

$$\mathbf{Tr}\left(\mathbf{S}^{-1}\right)\leq 1 \quad\Rightarrow\quad \gamma=\frac{1}{\mathbf{Tr}\left(\left(\mathbf{W}\mathbf{W}^T\right)^{\frac{1}{2}}\right)}$$

According to the above, a redundancy matrix $\mathbf{S}$ can be got as

$$\begin{aligned}\mathbf{S}^{-1}&=\frac{\left(\mathbf{W}\mathbf{W}^T\right)^{\frac{1}{2}}}{\mathbf{Tr}\left(\mathbf{W}\mathbf{W}^T\right)^{\frac{1}{2}}}\\&\Rightarrow\mathbf{S}=\mathbf{Tr}\left(\left(\mathbf{W}\mathbf{W}^T\right)^{\frac{1}{2}}\right)\left(\mathbf{W}\mathbf{W}^T\right)^{-\frac{1}{2}}\end{aligned} \tag{21}$$

3) **Update** $\mathbf{W}$ with $\mathbf{V}$, $\mathbf{S}$, λ fixed

By taking the derivative with respect to $\mathbf{W}$ and setting it to zero, we can obtain

$$\mathbf{W} = \left(\mathbf{X}\mathbf{L}^{(\lambda)}\mathbf{X}^T + \mathbf{A} + \mathbf{S} + \mu\mathbf{I}\right)\left(\mathbf{V} - \frac{\Lambda}{\mu}\right) \tag{22}$$

where $\mathbf{I} \in \mathbb{R}^{d\times d}$ is the identity matrix.

4) **Update V** with $\mathbf{W}$, $\mathbf{S}$, λ fixed

The subproblem can become,

$$\min_{\|\mathbf{V}\|_{2,0}=k} \left\|\mathbf{W} - \mathbf{V} + \frac{\Lambda}{\mu}\right\|_F^2 \tag{23}$$

which can be tackled by Algorithm 2.

Algorithm 2 A Solution of Problem (23)

Input: $\mathbf{W}$, Λ, μ and k.

Process:

1: Compute $\widetilde{\mathbf{W}} = \mathbf{W} + \frac{1}{\mu}\Lambda$.
2: $\mathbf{w} = diag(\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^T)$.
3: Sort $\mathbf{w}$, find out the indices vector $\mathbf{q} = [q_1, q_2, \ldots, q_k]^T$ corresponding to top k sorted entries.
4: Set i-th row of $\widetilde{\mathbf{W}}$ to $\mathbf{V}$ if $i \in \mathbf{q}$; set zero row of $\mathbf{0}^T \in \mathbb{R}^{1\times m}$ if $i \notin \mathbf{q}$.

Output: $\mathbf{V}$.

Algorithm 3 SSD-RM Method

Input: Training data $\mathbf{X}$, training labels $\mathbf{Y}$, The number of feature selected k, $\mathbf{L}_b$, $\mathbf{L}_w$ and Redundancy matrix $\mathbf{A}$

Output: Projection matrix $\mathbf{W} \in \mathbb{R}^{d\times m}$

Initialization:

1: $\mathbf{W} = \mathbf{W}_0$
2: $\Lambda \in \mathbf{0}^{d\times m}$

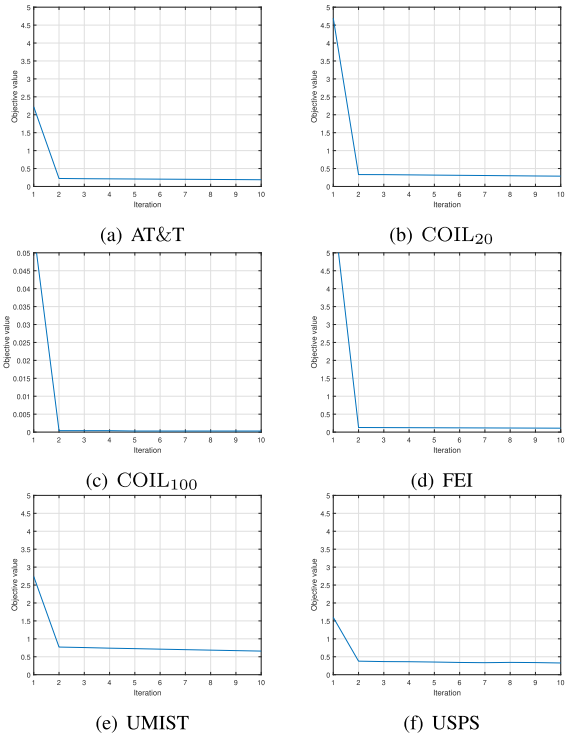
Process:

1: **repeat**
2: Update $\lambda \leftarrow \frac{\mathbf{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}_b\mathbf{X}^T\mathbf{W})}{2\mathbf{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}_w\mathbf{X}^T\mathbf{W})}$
3: Update $\mathbf{S} \leftarrow \mathbf{Tr}\left(\left(\mathbf{W}\mathbf{W}^T\right)^{\frac{1}{2}}\right)\left(\mathbf{W}\mathbf{W}^T\right)^{-\frac{1}{2}}$
4: Update $\mathbf{L}^{(\lambda)} \leftarrow \lambda\mathbf{L}_w - \mathbf{L}_b$, $\mathbf{G} = \mathbf{X}\mathbf{L}^{(\lambda)}\mathbf{X}^T + \mathbf{A}$
5: Update $\mathbf{W} \leftarrow (\mathbf{G} + \mathbf{S} + \mu\mathbf{I})^{-1}\left(\mathbf{V} - \frac{\Lambda}{\mu}\right)$
6: Update $\mathbf{V}$ by Alg.2.
7: Update Λ by $\Lambda + \mu\,(\mathbf{W} - \mathbf{V})$
8: Update μ by $\rho\mu$
9: **until** Converges

Therefore, we can iteratively update λ, $\mathbf{S}$, $\mathbf{W}$, $\mathbf{V}$ and our algorithm is summarized in Algorithm 3.

V. EXPERIMENT

To demonstrate the effectiveness and superiority of our method, we compare our method (called as SSD-RM) to the several other state-of-the-art feature selection methods:



(a) AT&T (b) COIL_{20} (c) COIL_{100} (d) FEI (e) UMIST (f) USPS

FIGURE 1. Convergence of the proposed SSD-RM on six data sets.

T-test [31], mRMR [32], TRC-FS [7], RALM-FS [25], CR-FS [33] SSD-FS [14] and SDFS-ARM [29]. Besides, we will use 50% of the input data as training set and the remainder as test set in the experiment. Moreover, two classifiers, such as the conventional K-NN [34] with Euclidean distance metric and linear SVM [35], [36], are used to computing the classification accuracies on test set. Meanwhile, the parameter C of linear SVM is set from {0.001, 0.01, 0.1, 1, 10, 100}, using 10-fold cross-validation. The parameter of K-NN is set as K=1. In addition, The formula for calculating classification accuracy is

$$\textbf{Classification\_Acc} = \frac{\textbf{Num\_Correct}}{\textbf{Num\_Test}} \times \mathbf{100\%} \tag{24}$$

where **Num_Correct** and **Num_Test** denote correct numbers of classified samples and whole samples on the test set.

At the same time, we can adopt the following formula to computer the correlation (redundancy) among the features:

$$\text{Correlation } (C) = \frac{1}{n(n-1)} \sum_{f_i, f_j \in S, i \neq j} \mathbf{A}_{i,j} \tag{25}$$

where C is all the selected features, $\mathbf{A}$ is the redundancy matrix which can be computed by the square cosine similarity and n is the number of selected features in C.

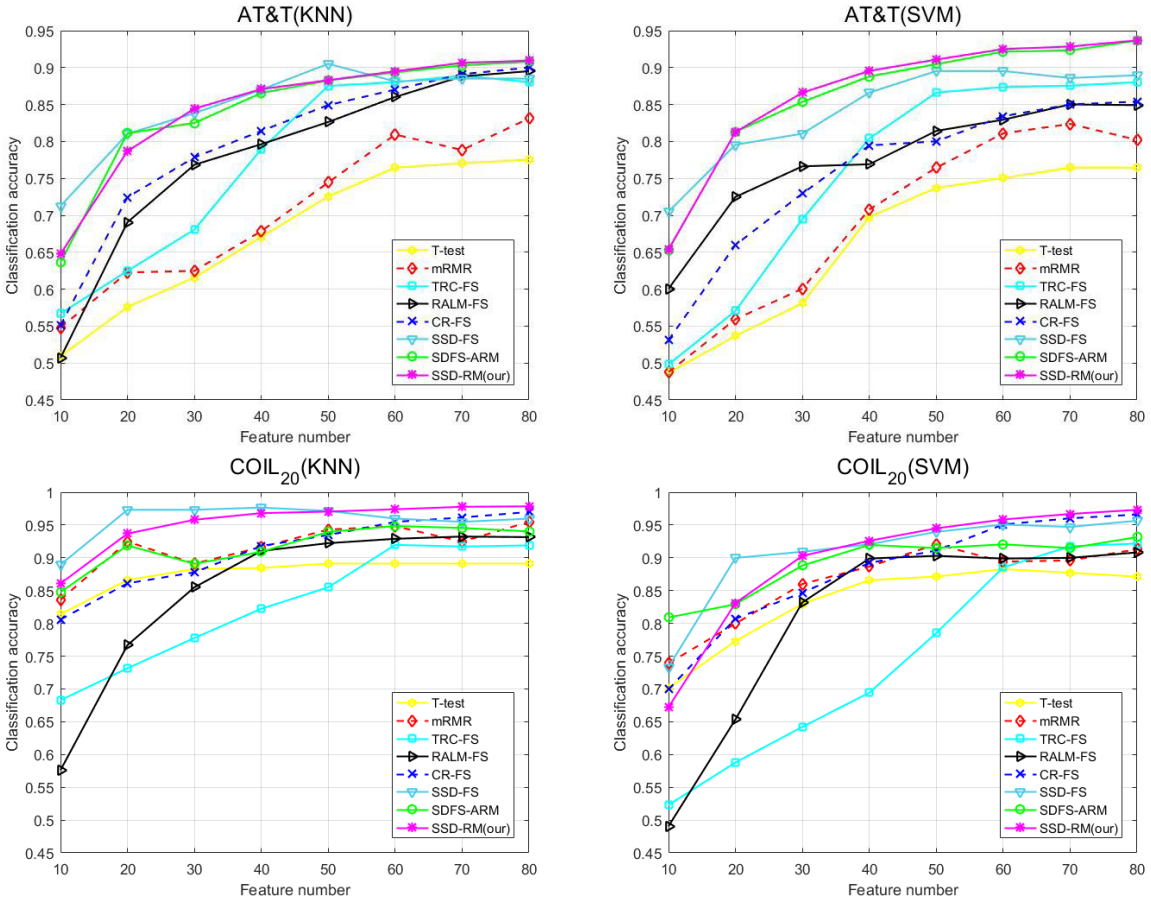


FIGURE 2. Comparison of the classification accuracy performed under two benchmark data sets.

TABLE 1. Datasets description.

Dataset	#Samples	#Classes	#Features
AT&T	400	40	1024
COIL_{20}	1440	20	1024
COIL_{100}	7200	100	1024
FEI	2800	200	1024
UMIST	575	20	1024
USPS	9298	10	256

A. DATA SETS DESCRIPTION

In this section, we use six real-world datasets, i.e., AT&T,[1] $COIL_{20}$,[2] $COIL_{100}$,[3] FEI,[4] $UMIST$[5] and $USPS$,[6] to verify the performance of our proposed feature selection method. Table 1 lists more details for each data set. Besides, we set the parameter $m = c$ for the fairness of the experiment.

B. CLASSIFICATION ACCURACY COMPARED WITH THE PREVIOUS METHODS

Finally, both K-NN classifier and SVM classifier are performed on six real-world datasets so as to obtain the classification accuracy of the feature selection methods. Some of the experimental data comes from the experimental results published in [14]. Bold face represents best result and the underlined represents second best result.

In Figure 1, the convergent curves of our SSD-RM method can be obtained on six data sets. In Table 2, we select top 80 features under 4 data sets: $COIL_{100}$, $UMIST$, $USPS$ and FEI, such that the average classification accuracy and standard deviation can be compared. In Figure. 2, the classification accuracies are compared by using increasing number of features on the datasets $AT\&T$ and $COIL_{20}$. According to Table 2 and Figure. 2, we can draw the following conclusions.

1) Table 2 shows best results can be obtained by our method in most data sets. Besides, although our method cannot get best results in several data sets, but suboptimal results may be obtained. Therefore, our method outperform other feature selection methods.

2) From Figure. 2, we can see that our method has better performance than other methods in the case of most feature numbers.

From the above conclusions, the proposed SSD-RM method can select the discriminative and non-redundant features by minimizing the redundancy between features. Therefore, our method has better classification performance.

[1] http://www.cl.cam.ac.uk/Research/DTG/attarchive/pub/data/att_faces.zip
[2] http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
[3] http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php
[4] http://fei.edu.br/ cet/facedatabase.html
[5] http://eeepro.shef.ac.uk/vie/face.tar.gz
[6] http://www-i6.informatik.rwth-aachen.de/ keysers/usps.html

TABLE 2. Classification accuracy (%) comparisons and standard deviation of the selected top 80 features on the dataset $COIL_{100}$, UMIST, USPS and FEI.

Dataset	$COIL_{100}$		UMIST	
Classification Accuracy	K-NN(%)	SVM(%)	K-NN(%)	SVM(%)
T-test [31]	77.05±2.34	77.86±2.44	87.08±2.03	88.29±1.51
mRMR [32]	81.43±2.11	83.32±1.58	88.75±1.23	88.92±1.38
TRC-FS [7]	83.45±1.41	84.63±1.07	92.45±2.39	93.43±1.52
RALM-FS [25]	83.80±1.62	84.50±1.57	92.78±1.53	93.68±2.45
CR-FS [33]	84.48±2.22	85.03±2.00	92.50±1.57	94.05±1.11
SSD-FS [14]	85.26±2.02	85.11±2.18	93.66±1.86	94.88±1.08
SDFS-ARM [29]	85.13±1.53	86.54±1.24	**95.40±1.58**	94.73±0.97
SSD-RM(our)	**93.28±0.61**	**93.29±0.75**	94.44±1.95	**94.99±1.65**
Dataset	USPS		FEI	
Classification Accuracy	K-NN(%)	SVM(%)	K-NN(%)	SVM(%)
T-test [31]	92.33±0.30	93.06±0.37	71.11±1.24	72.23±2.05
mRMR [32]	94.75±0.33	94.01±0.31	73.42±1.98	74.05±1.26
TRC-FS [7]	92.62±0.83	93.34±0.59	73.36±1.14	74.56±2.23
RALM-FS [25]	94.68±0.77	94.47±0.51	61.36±1.04	69.70±1.12
CR-FS [33]	91.06±1.97	92.67±1.12	74.64±1.61	75.78±2.33
SSD-FS [14]	75.29±0.93	77.58±2.04	**75.46±1.73**	75.96±1.82
SDFS-ARM [29]	93.08±1.12	93.92±0.95	73.40±1.89	74.13±1.76
SSD-RM(our)	**95.32±0.67**	**94.68±0.38**	72.25±2.58	**77.24±1.75**

TABLE 3. Classification accuracy and correlation of top 80 features via K-NN classifier.

	datasets	SSD-RM	without A	without S
Accuracy	UMIST	**0.9444**	0.9427	0.8459
	USPS	**0.9532**	0.9526	0.8407
	$COIL_{100}$	**0.9328**	0.9319	0.8873
	$COIL_{20}$	**0.9788**	0.9780	0.8175
	FEI	**0.7225**	0.7199	0.2010
	AT&T	**0.9088**	0.9083	0.8663
Correlation	UMIST	0.0962	**0.0959**	0.4386
	USPS	0.0919	**0.0916**	0.1540
	$COIL_{100}$	0.1120	0.1133	**0.0967**
	$COIL_{20}$	**0.0918**	0.0927	0.2456
	FEI	0.1060	**0.1052**	1.0078
	AT&T	0.0793	**0.0777**	0.2356

TABLE 4. Classification accuracy and correlation of top 80 features via SVM classifier.

	datasets	SSD-RM	without A	without S
Accuracy	UMIST	**0.9499**	0.9451	0.6531
	USPS	**0.9468**	0.9458	0.8535
	$COIL_{100}$	**0.9329**	0.9324	0.8429
	$COIL_{20}$	**0.9739**	0.9735	0.7224
	FEI	**0.7724**	0.7667	0.5997
	AT&T	**0.9378**	0.9365	0.8846
Correlation	UMIST	0.0954	**0.0947**	0.4394
	USPS	**0.0915**	0.0925	0.1583
	$COIL_{100}$	0.1126	0.1123	**0.0964**
	$COIL_{20}$	**0.0925**	0.0931	0.2540
	FEI	0.1075	**0.1048**	1.0077
	AT&T	0.0788	**0.0783**	0.2383

C. CLASSIFICATION ACCURACY AND CORRELATION COMPARISONS WITHOUT THE MATRIX A OR WITHOUT THE MATRIX S

From TABLE 3 and TABLE 4, we can see the adaptive redundancy matrix $\mathbf{S}$ plays a vital role in our algorithm SSD-RM. Besides, the redundancy matrix $\mathbf{A}$ has a certain improvement in the performance of our algorithm. Besides, the correlation among the selected features is largely reduced by introducing the adaptive redundancy matrix $\mathbf{S}$ and the redundancy matrix $\mathbf{A}$. Therefore, we can obtain non-redundant discriminative features.

For this experimental results, we can analyze and explain from a theoretical perspective. Firstly, the redundancy $\mathbf{A}$ can only measure the initial correlation among the features. However, we treat the adaptive redundancy matrix $\mathbf{S}$ as a variable and use the adaptive redundancy matrix $\mathbf{S}$ to measure correlation among the features in each iteration of the experiment. Therefore, the above analysis confirms the experimental results.

D. COMPUTATIONAL COMPLEXITY

In this section, for validating the superiority of our proposed method, we analyze the computational complexity of the proposed SSD-RM method. From Algorithm 3, we can see the computational complexity of SSD-RM algorithm can be ascribed to the calculation of the matrix $\mathbf{S}$ and the matrix $\mathbf{W}$. Therefore, the computational complexity of SSD-RM method is $\mathcal{O}(dmn + d^2n)$, where d, n, m denote the numbers of all features, samples and selectable features, respectively $(1 \leq m < d)$.

VI. CONCLUSION

In this paper, we apply self-weighted linear discriminant analysis (SLDA) method and introduce $l_{2,0}$-norm to SLDA problem. Because of directly solving the original $l_{2,0}$-norm constrained problem, we can avoid spending some time to adjust the regularization parameters whose regularization parameter has a clear meaning, which is the number of selected features. As we all known, the problem of feature selection needs to be considered from two aspects: the discriminant ability of features and the correlation among features. For the feature redundancy problem, we introduce the redundancy matrix $\mathbf{A}$ in the AGRM framework to SLDA problem. Meanwhile, we introduce the adaptive redundancy matrix $\mathbf{S}$ and treat the redundancy matrix $\mathbf{S}$ as an optimizing variable. Finally, we propose an efficient algorithm to address the constrained optimization problem, which is based on the augmented Lagrangian method, such that the global optimal solution can be obtained. In the future, we pay attention to apply our algorithm to some computer vision tasks, such as image segmentation [37], object detection [38] and so on.

REFERENCES

[1] 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.

[2] F. Nie, W. Zhu, X. Li, "Unsupervised feature selection with structured graph optimization," in *Proc. 30th AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2016, pp. 1302–1308.

[3] 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*, 2010, pp. 333–342.

[4] X. Chen, G. Yuan, F. Nie, and Z. Ming, "Semi-supervised feature selection via sparse rescaled linear square regression," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 1, pp. 165–176, Jan. 2020.

[5] Z. Xu, I. King, M. R.-T. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *IEEE Trans. Neural Netw.*, vol. 21, no. 7, pp. 1033–1047, Jul. 2010.

[6] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norm minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.

[7] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection," in *Proc. Conf. Artif. Intell. (AAAI)*, vol. 2, 2008, pp. 671–676.

[8] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria," *Ann. Math. Artif. Intell.*, vol. 41, no. 1, pp. 77–93, 2004.

[9] K. Kira, L. A. Rendell, "A practical approach to feature selection," in *Machine Learning Proceedings*. Amsterdam, The Netherlands: Elsevier, 1992, pp. 249–256.

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

[11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.

[12] J. R. Quinlan, *C4. 5: Programs for Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2014.

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

[14] R. Zhang, F. Nie, and X. Li, "Self-weighted supervised discriminative feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3913–3918, Aug. 2018.

[15] X. Wu, J. Liu, H. Wang, B. Hu, and F. Nie, "Supervised feature selection with orthogonal regression and feature weighting," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 14, 2020, doi: 10.1109/TNNLS.2020.2991336.

[16] H. Zhang, J. Wang, Z. Sun, J. M. Zurada, and N. R. Pal, "Feature selection for neural networks using group lasso regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 4, pp. 659–673, Apr. 2020.

[17] W. Liu and I. W. Tsang, "Making decision trees feasible in ultrahigh feature and label dimensions," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 2814–2849, 2017.

[18] W. Liu and I. Tsang, "Sparse perceptron decision tree for millions of dimensions," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1881–1887.

[19] W. Liu, X. Shen, and I. Tsang, "Sparse embedded k-means clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3321–3329.

[20] R. Tibshirani, I. Johnstone, T. Hastie, and B. Efron, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, Apr. 2004.

[21] J. Liu, J. Chen, and J. Ye, "Large-scale sparse logistic regression," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining- KDD*, 2009, pp. 547–556.

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

[23] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "$\ell_{2,0}$-norm regularized discriminative feature selection for unsupervised learning," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1589–1594.

[24] H. Yan, J. Yang, and J. Yang, "Robust joint feature weights learning framework," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 5, pp. 1327–1339, May 2016.

[25] X. Cai, F. Nie, and H. Huang, "Exact top-k feature selection via $\ell_{2,0}$-norm constraint," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1240–1246.

[26] T. Pang, N. Feiping, H. Junwei, and L. Xuelong, "Efficient feature selection via $\ell_{2,0}$-norm constrained sparse regression," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 5, pp. 880–893, May 2019.

[27] Z. Wang, F. Nie, L. Tian, R. Wang, and X. Li, "Discriminative feature selection via a structured sparse subspace learning module," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 3009–3015.

[28] F. Nie, S. Yang, R. Zhang, and X. Li, "A general framework for auto-weighted feature selection via global redundancy minimization," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2428–2438, May 2019.

[29] T. Wu, Y. Zhou, R. Zhang, Y. Xiao, and F. Nie, "Self-weighted discriminative feature selection via adaptive redundancy minimization," *Neurocomputing*, vol. 275, pp. 2824–2830, Jan. 2018.

[30] S. Boyd, L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.

[31] D. C. Montgomery and D. C. Hubele, *Engineering Statistics*. Hoboken, NJ, USA: Wiley, 2007.

[32] 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.

[33] R. He, T. Tan, L. Wang, and W.-S. Zheng, "$\ell_{2,0}$ regularized correntropy for robust feature selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2504–2511.

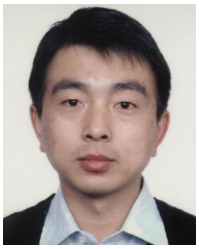
[34] O. Kramer, *Dimensionality Reduction With Unsupervised Nearest Neighbors*. Berlin, Germany: Springer, 2013.

[35] C. Cortes and V. Vapnik, "Support vector machine," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[36] R. Hu, X. Zhu, Y. Zhu, and J. Gan, "Robust SVM with adaptive graph learning," *World Wide Web*, vol. 23, no. 3, pp. 1945–1968, May 2020.

[37] M. M. Cheng, Y. Liu, Q. Hou, J. Bian, P. Torr, S. M. Hu, and Z. Tu, "HFS: Hierarchical feature selection for efficient image segmentation," *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands: Springer, 2016, pp. 867–882.

[38] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.



HAIHONG YU was born in 1975. He received the Ph.D. degree from Jilin University. He is currently a Professor with Jilin University. His main research interests include decision support systems and machine learning. He is a member of CCF.

LIANGLIANG ZHANG is currently pursuing the master's degree in software engineering with Jilin University. His research interests include feature selection and sparse learning.

ZHANSHAN LI is currently a Professor of computer science with Jilin University. His main research interests include constraint reasoning and machine learning.