Received: 15 July 2020 | Revised: 9 October 2020 | Accepted: 3 November 2020

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12018

ORIGINAL RESEARCH PAPER

Low-rank constrained weighted discriminative regression for multi-view feature learning

Chao Zhang | Huaxiong Li

Department of Control and Systems Engineering,
Nanjing University, Nanjing, 210093, China

Correspondence

Huaxiong Li, Department of Control and Systems Engineering, Nanjing University, Nanjing, 210093, China.
Email: huaxiongli@nju.edu.cn

Funding information

National Nature Science Foundation of China, Grant/Award Numbers: 71671086, 71732003

Abstract

In recent years, multi-view learning has attracted much attention in the fields of data mining, knowledge discovery and machine learning, and been widely used in classification, clustering and information retrieval, and so forth. A new supervised feature learning method for multi-view data, called low-rank constrained weighted discriminative regression (LWDR), is proposed. Different from previous methods handling each view separately, LWDR learns a discriminative projection matrix by fully exploiting the complementary information among all views from a unified perspective. Based on least squares regression model, the high-dimensional multi-view data is mapped into a common subspace, in which different views have different weights in projection. The weights are adaptively updated to estimate the roles of all views. To improve the intra-class similarity of learned features, a low-rank constraint is designed and imposed on the multi-view features of each class, which improves the feature discrimination. An iterative optimization algorithm is designed to solve the LWDR model efficiently. Experiments on four popular datasets, including Handwritten, Caltech101, PIE and AwA, demonstrate the effectiveness of the proposed method.

1 | INTRODUCTION

Owing to the rapid growth of multimedia data, multi-view learning aroused much interests of researchers from data mining, knowledge discovery and machine learning areas in recent years [1–8]. In real world, one object can be described by different kinds of data or from different views. For example, a news can be expressed by texts, audios and videos. One person can be identified by the face, fingerprint and DNA information. Although these features may be heterogeneous and very different, they naturally reflect some inherent structures or characteristics of the object. Compared with single view data, multi-view data contains more underlying information. How to effectively and efficiently exploit the correlative yet complementary information among diverse views is an important research topic for multi-view learning [9].

Many researchers tried to develop effective information fusion techniques for multi-view learning, including unsupervised [9], semi-supervised [10] and supervised [11, 12]. Some researchers proposed multi-view learning methods by co-training [13, 14] and co-regularization [15, 16]. However, these methods neglect the problem caused by the high dimension of multi-view data. Canonical correlation analysis (CCA) is a classical unsupervised multi-view subspace learning method, which seeks a low dimensional feature space by maximizing the correlation between different views [9]. However, CCA can only deal with two views which limits its further application on more complex data. Luo et al. proposed a tensor CCA method, which extended original CCA for multiple views [17]. To make use of the label information for better classification performance, Kan et al. proposed a multi-view discriminant analysis (MvDA) method by maximizing the inter-class distance and minimizing the intra-class distance from both inter-view and intra-view [11]. MvDA can be regarded as the extension of linear discriminant analysis (LDA) on multi-view data. These methods mentioned above can be categorized into subspace learning based methods, which assume that different views can be generated from a common latent feature subspace.

Regression-based method is one of the most popular methods in machine learning [18–22], and it provides another effective and efficient way for multi-view feature learning [23]. Specifically, regression-based methods seek a linear mapping by transforming data to fit the label matrix. Zheng et al. extended low-rank regression model for single view data to multi-view fully low-rank regression (FLR), in which multiple projection matrices were learned and the final classification was performed by majority voting [23]. However, FLR does not consider the differences among all views. Yang et al. proposed an adaptive-weighting discriminative regression (ADR) by learning a unified transform matrix for multi-view classification [24]. ADR introduces an adjustment matrix to enlarge the distance between different classes which improves the model robustness to some extent. However, the intra-class compactness is destroyed in ADR, which is also important for pattern analysis. In [25], the authors incorporated feature selection into linear regression model by $l_{2,1}$-norm regularization. Wen et al. adopted a graph-regularized matrix factorization model to handle the multi-view clustering problem with incomplete views [7].

In this article, we propose a multi-view low-rank weighted discriminative regression (LWDR) method for feature learning. LWDR learns a common feature subspace across all views. Adaptive weights learning mechanism is adopted to automatically learn different views and the important views containing more discriminative information is enforced to contribute more to subspace learning. To improve the intra-class similarity, a class-wise low-rank constraint is imposed on the multi-view features. Besides, a flexible error term based on $l_{2,1}$-norm is introduced to relax the label matrix. Experiments on four datasets demonstrate that LWDR outperforms previous single view feature learning methods and related multi-view learning methods.

The rest of this paperarticle is organized as follows. In Section 2, we briefly review the related works about linear regression methods for multi-view learning. Section 3 introduces the formulation of our proposed method and the optimization algorithm in detail. Section 4 reports the experimental results and analysis. Section 5 concludes this article.

2 | RELATED WORKS

For convenience, we first present the main notations used in this paper. matrices and vectors are written in boldface uppercase and boldface lowercase, respectively. $\{\mathbf{X}_k\}_{k=1}^{v}$ denotes a multi-view dataset with v views. $\mathbf{X}_k \in \mathbb{R}^{m_k \times n}$ is the data matrix of the k-th view, where m_k is the feature dimensionality of view k and n is the number of training samples. c is the number of classes. $\mathbf{Y} \in \{0,1\}^{c \times n}$ is the label matrix, and $Y_{ij} = 1$ if the j-th training sample belongs to class i and otherwise 0. $\mathbf{W}_k \in \mathbb{R}^{m_k \times c}$ is the projection matrix of the k-th view. For matrix $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{p \times q}$, its $l_{2,1}$-norm, Frobenius-norm, and nuclear norm are defined as $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{q} A_{ij}^2}$, $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{q} A_{ij}^2}$, $\|\mathbf{A}\|_* = \sum_{i=1} \delta_i(\mathbf{A})$, respectively, where $\delta_i(\mathbf{A})$ is the i-th singular value of $\mathbf{A}$. $\mathbf{I}$ is an identity matrix and $\mathbf{1}_n$ is an n-dimensional vector with all elements being 1. $\odot$ is the Hadamard multiplication.

For single view data $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $\mathbf{Y}$, the basic linear regression solves the following problem to find the optimal projection matrix $\mathbf{W}$.

$$\min_{\mathbf{W}} \frac{1}{2} \|\mathbf{W}^T \mathbf{X} - \mathbf{Y}\|_F^2 + \alpha \phi(\mathbf{W}), \tag{1}$$

where $\alpha > 0$ is a balance parameter and $\phi(\mathbf{W})$ is the regularizer. $l_{2,1}$-norm, Frobenius-norm and nuclear norm are usually used in $\phi(\mathbf{W})$ for learned features with different properties. When $\phi(\mathbf{W}) = \|\mathbf{W}\|_F^2$, Equation (1) has a closed-form solution $\mathbf{W}^* = (\mathbf{X}\mathbf{X}^T + 2\alpha\mathbf{I})^{-1}\mathbf{X}\mathbf{Y}^T$, which is very efficient in real application. However, Equation (1) cannot be directly applied on multi-view data. Some researchers extended it for multi-view data by learning multiple projection matrices as follows [23]:

$$\min_{\mathbf{W}} \frac{1}{2} \| \sum_{k=1}^{v} \mathbf{W}_k^T \mathbf{X}_k - \mathbf{Y} \|_F^2 + \alpha \sum_{k=1}^{v} \phi(\mathbf{W}_k), \tag{2}$$

where $\mathbf{W}_k$ is the projection matrix of the k-th view. It can be observed that Equation (2) is equivalent to simply concatenating multi-view features into one single view feature and performing traditional linear regression on it. Such operation is not physically meaningful and treats all views equally. However, different views usually have different characteristics and contribute differently to pattern analysis. For example, to identify a person, the face appearance is more important than voice. Thus, adaptive weights learning strategy is introduced into multi-view linear regression model by weighting different views, which can be generally described as follows.

$$\begin{aligned} \min_{\mathbf{W}} \quad & \frac{1}{2} \| \sum_{k=1}^{v} \delta_k \mathbf{W}_k^T \mathbf{X}_k - \mathbf{Y} \|_F^2 + \alpha \sum_{k=1}^{v} \phi(\mathbf{W}_k) \\ \text{s.t.} \quad & \delta_k > 0, \sum_{k=1}^{v} \delta_k = 1. \end{aligned} \tag{3}$$

where δ_k is the weight of the k-th view. The first term and constraints force the model to learn optimal weight for each view automatically. However, in Equation (3), the latent multi-view features are directly regressed to approximate the binary label matrix $\mathbf{Y}$, which may be too strict and is not appropriate as the regression target. To improve the model robustness, in [24], the authors relaxed the label matrix and proposed the following auto-weighted discriminative regression model for multi-view classification (ADR):

$$\begin{aligned} \min & \frac{1}{2} \| \sum_{k=1}^{v} \sqrt{\delta_k} \mathbf{W}_k^T \mathbf{X}_k - \mathbf{b}\mathbf{1}_n^T - (\mathbf{Y} + \mathbf{Y} \odot \mathbf{M}) \|_F^2 \\ & + \alpha \sum_{k=1}^{v} \|\mathbf{W}_k\|_F^2, \\ & \text{s.t. } \mathbf{M} \geq 0, \delta_k > 0, \sum_{k=1}^{v} \delta_k = 1. \end{aligned} \tag{4}$$

where $\mathbf{b} \in \mathbb{R}^c$ is a bias vector and $\mathbf{M}$ is a non-negative matrix. By introducing the adjustment matrix $\mathbf{M}$, the one-zero label vector $\mathbf{y} = [1,0,\ldots,0]^{\mathrm{T}}$ is relaxed to

$[1+M_{ij},0,\ldots,0]^{\mathrm{T}}(M_{ij}\geq 0)$, where M_{ij} is the corresponding element in $\mathbf{M}$. By introducing non-negative matrix $\mathbf{M}$, Equation (4) uses the ϵ-draggings technique to enlarge the distances between true and false classes and improve the model robustness. Such label relaxation strategy is widely used in other regression-based methods [26, 27].

3 | PROPOSED METHOD

In this section, we present our proposed LWDR method and the optimization algorithm in detail, then we will make some discussions on the computational complexity of LWDR.

3.1 | Formulation

From Section 2, although the ϵ-dragging technique used in ADR [24] enlarges the distance between different classes or inter-class separability to some extent, the same class may have different labels after relaxation (i.e. $\mathbf{Y}+\mathbf{Y}\odot\mathbf{M}$) because of the dynamic of $\mathbf{M}$. Thus, the intra-class similarity cannot be guaranteed in ADR and two samples from same class may be projected distant from each other. Both inter-class separability and intra-class similarity are important for good classification performance. In low-rank representation (LRR) [28–30], a set of instances are generally drawn from a union of multiple subspaces, and the instances of each subspace are regarded from the same class, which illustrates that the instances from the same class should locate in the same subspace and the data matrix of each class should be low-rank. Inspired by this issue, we propose the following low-rank constrained adaptive weighted discriminative regression (LWDR) model to improve the intra-class similarity of learned multi-view features:

$$\begin{aligned}\min\ &\frac{1}{2}\Big\|\sum_{k=1}^{v}\sqrt{\delta_k}\mathbf{W}_k^T\mathbf{X}_k-\mathbf{H}\Big\|_F^2+\frac{1}{2}\|\mathbf{H}+\mathbf{E}-\mathbf{Y}\|_F^2\\ &+\alpha\sum_{i=1}^{c}\|\mathbf{H}_i\|_*+\beta\sum_{k=1}^{v}\|\mathbf{W}_k\|_F^2+\gamma\|\mathbf{E}\|_{2,1},\\ &\text{s.t.}\ \ \delta_k>0,\ \sum_{k=1}^{v}\delta_k=1.\end{aligned}\tag{5}$$

where α, β and γ are balance parameters, $\mathbf{E}$ is an error matrix, $\mathbf{H}$ is the to-be-learned feature matrix, and $\mathbf{H}_i$ denotes the features of the i-th class. The first term in Equation (5) learns the multi-view features approximated by matrix $\mathbf{H}$. The second term utilizes the label matrix to supervise the feature learning. The third term $\sum_{i=i}^{c}\|\mathbf{H}_i\|_*$ forces the features of same class to be similar, which are considered to be low-rank. The last two terms are regularizers avoiding overfitting. Error matrix $\mathbf{E}$ is constrained by $l_{2,1}$-norm to compensate the regression errors and it is flexible in learning transform matrices $\{\mathbf{W}_k\}_{k=1}^{v}$. The overall framework of our LWDR is shown in Figure 1. Equation (5) involves multiple variables and cannot be solved directly. In the next section, we will give an iterative algorithm to solve it efficiently.

3.2 | Optimization

To make the variables separable in Equation (5), we introduce an auxiliary variable $\mathbf{T}$ and rewrite the original problem as follows:

$$\begin{aligned}\min\ &\frac{1}{2}\Big\|\sum_{k=1}^{v}\widetilde{\mathbf{W}}_k^T\mathbf{X}_k-\mathbf{H}\Big\|_F^2+\frac{1}{2}\|\mathbf{H}+\mathbf{E}-\mathbf{Y}\|_F^2\\ &+\alpha\sum_{i=1}^{c}\|\mathbf{T}_i\|_*+\beta\sum_{k=1}^{v}\frac{1}{\delta_k}\|\widetilde{\mathbf{W}}_k\|_F^2+\gamma\|\mathbf{E}\|_{2,1},\\ &\text{s.t.}\ \ \mathbf{T}=\mathbf{H},\ \delta_k>0,\ \sum_{k=1}^{v}\delta_k=1,\end{aligned}\tag{6}$$

where $\widetilde{\mathbf{W}}_k=\sqrt{\delta_k}\mathbf{W}_k$. By some simple manipulation, the above problem can be equivalently expressed as:

$$\begin{aligned}\min\ &\frac{1}{2}\|\widetilde{\mathbf{W}}^T\mathbf{X}-\mathbf{H}\|_F^2+\frac{1}{2}\|\mathbf{H}+\mathbf{E}-\mathbf{Y}\|_F^2\\ &+\alpha\sum_{i=1}^{c}\|\mathbf{T}_i\|_*+\beta\sum_{k=1}^{v}\frac{1}{\delta_k}\|\widetilde{\mathbf{W}}_k\|_F^2+\gamma\|\mathbf{E}\|_{2,1},\\ &\text{s.t.}\ \ \mathbf{T}=\mathbf{H},\ \delta_k>0,\ \sum_{k=1}^{v}\delta_k=1.\end{aligned}\tag{7}$$

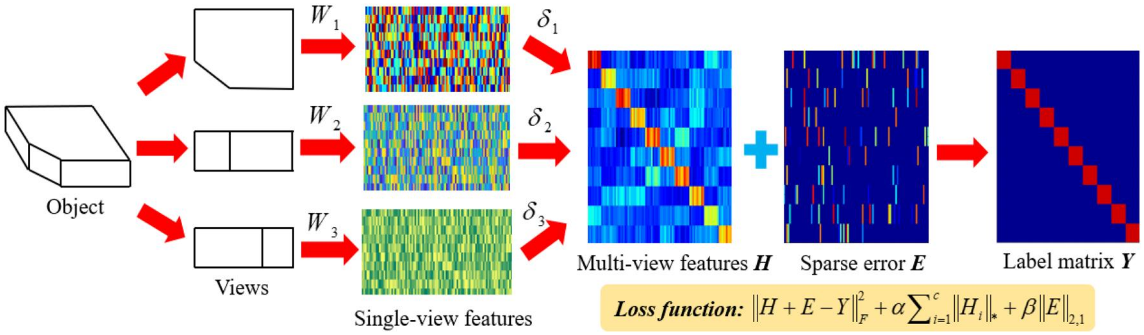

FIGURE 1 The overall framework of LWDR model for multi-view feature learning. LWDR adaptively learns a projection matrix and a weight for each view. Then the multi-view features $\mathbf{H}$ can be obtained by integrating the single-view features using learned weights. Label matrix $\mathbf{Y}$ is used as regression target to guide the feature learning in a supervised way, and a sparse error matrix $\mathbf{E}$ is introduced to compensate the regression error. To improve the intra-class compactness of multi-view features, a class-wise low-rank constraint on multi-view features is incorporated into LWDR

where $\widetilde{\mathbf{W}} = [\widetilde{\mathbf{W}}_1; \widetilde{\mathbf{W}}_2; \ldots; \widetilde{\mathbf{W}}_v], \mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2; \ldots; \mathbf{X}_v]$. To solve, Equation (7) is equivalent to minimize its augmented Lagrange function $\mathcal{L}_\theta$ defined as:

$$\begin{aligned}\mathcal{L}_\theta = & \frac{1}{2}\|\widetilde{\mathbf{W}}^T\mathbf{X} - \mathbf{H}\|_F^2 + \frac{1}{2}\|\mathbf{H} + \mathbf{E} - \mathbf{Y}\|_F^2 \\ & + \alpha\sum_{i=1}^{c}\|\mathbf{T}_i\|_* + \beta\sum_{k=1}^{v}\frac{1}{\delta_k}\|\widetilde{\mathbf{W}}_k\|_F^2 + \gamma\|\mathbf{E}\|_{2,1}, \\ & + \frac{\theta}{2}\|\mathbf{T} - \mathbf{H} + \frac{1}{\theta}\mathbf{Z}\|_F^2, \end{aligned} \tag{8}$$

where $\theta > 0$ is a penalty factor and $\mathbf{Z}$ is the augmented Lagrange multiplier. We adopt the iterative strategy to minimize it, in which a sequence of sub-problems with respect to each unknown variable are solved respectively [31–33]. In specific, it contains following six steps in each iteration.

(1) Update $\widetilde{\mathbf{W}}$: We first reorganize $\{\delta_k\}_{k=1}^{v}$ as δ and rewrite the sub-problem w.r.t $\widetilde{\mathbf{W}}$ as follows:

$$\min_{\widetilde{\mathbf{W}}} \frac{1}{2}\|\widetilde{\mathbf{W}}^T\mathbf{X} - \mathbf{H}\|_F^2 + \frac{\beta}{\delta}\|\widetilde{\mathbf{W}}\|_F^2. \tag{9}$$

This is a typical least squares regression problem. By setting its derivative w.r.t. $\mathbf{W}$ to zero, we can obtain its optimal solution:

$$\widetilde{\mathbf{W}}^* = (2\beta\mathbf{I}/\delta + \mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{H}^T. \tag{10}$$

Since δ is a set of $\{\delta_k\}_{k=1}^{v}$, the final solution is:

$$\widetilde{\mathbf{W}}^* = (2\beta\Delta + \mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{H}^T, \tag{11}$$

where $\Delta = \mathrm{diag}(1/\delta_1, \ldots, 1/\delta_1, 1/\delta_2, \ldots, 1/\delta_v)$.

(2) Update $\mathbf{H}$ by solving the following minimization problem:

$$\begin{aligned}\min_{\mathbf{H}} \frac{1}{2}\|\widetilde{\mathbf{W}}^T\mathbf{X} - \mathbf{H}\|_F^2 \quad & + \frac{1}{2}\|\mathbf{H} + \mathbf{E} - \mathbf{Y}\|_F^2 \\ & + \frac{\theta}{2}\|\mathbf{T} - \mathbf{H} + \frac{1}{\theta}\mathbf{Z}\|_F^2,\end{aligned} \tag{12}$$

With the same strategy in optimizing $\widetilde{\mathbf{W}}$ and we can obtain the solution to Equation (12):

$$\mathbf{H}^* = ((2+\theta)\mathbf{I})^{-1}(\widetilde{\mathbf{W}}^T\mathbf{X} - \mathbf{E} + \mathbf{Y} + \theta\mathbf{T} + \mathbf{Z}). \tag{13}$$

(3) Update $\mathbf{E}$ by solving the following problem:

$$\min_{\mathbf{E}} \gamma\|\mathbf{E}\|_{2,1} + \frac{1}{2}\|\mathbf{E} + \mathbf{H} - \mathbf{Y}\|_F^2. \tag{14}$$

Equation (14) can be solved by the following theorem.

Theorem 1 [34] *Given* $\mathbf{Q} \in \mathbb{R}^{p\times q}$, *the optimal solution* $\boldsymbol{A}^*$ *of*

$$\min_{\mathbf{A}} \lambda\|\mathbf{A}\|_{2,1} + \frac{1}{2}\|\mathbf{A} - \mathbf{Q}\|_F^2,$$

is given by $G_\lambda(\mathbf{Q})$, *and* $G_\lambda(\mathbf{Q})$ *is the following operator:*

$$\mathbf{A}^*_{i,:} = G_\lambda(\mathbf{Q}_{i,:}) = \begin{cases} \dfrac{\|\mathbf{Q}_{i,:}\|_2 - \lambda}{\|\mathbf{Q}_{i,:}\|_2}\mathbf{Q}_{i,:}, & \text{if } \|\mathbf{Q}_{i,:}\|_2 > \lambda \\ 0, & \text{otherwise.} \end{cases}$$

where $\mathbf{Q}_i$ *is the* i-th row of matrix $\mathbf{Q}$.

According to Theorem 1, we can get the solution to Equation (14):

$$\mathbf{E}^* = G_\gamma(\mathbf{Y} - \mathbf{H}). \tag{15}$$

(4) Update $\mathbf{T}$ with other variables fixed by solving the following optimization problem:

$$\min_{\mathbf{T}} \alpha\sum_{i=1}^{c}\|\mathbf{T}_i\|_* + \frac{\theta}{2}\|\mathbf{T} - \mathbf{H} + \frac{1}{\theta}\mathbf{Z}\|_F^2. \tag{16}$$

Algorithm 1 Iterative algorithm for solving LWDR

```
Input: Multi-view data $\{\mathbf{X}_k\}_{k=1}^{V}$, parameters $\alpha$, $\beta$,
$\gamma$.
Output: Projection matrix $\widetilde{\mathbf{W}}$.
Initialization: $\mathbf{H} = \mathbf{Y}$, $\mathbf{E} = \mathbf{W} = \mathbf{Z} = \mathbf{0}$,
$\theta = 10^{-3}$, $\theta_{max} = 10^5$, $\epsilon = 10^{-6}$.
$t = 0$.
while not converged do.
    Update $\widetilde{\mathbf{W}}$ by Equation (11);
    Update $\mathbf{H}$ by Equation (13);
    Update $\mathbf{E}$ by Equation (15);
    Update $\mathbf{T}_i$ by Equation (18) and update
    $\mathbf{T} = [\mathbf{T}_1, \ldots, \mathbf{T}_c]$;
    Update $\{\delta_k\}_{k=1}^{V}$ by Equation (21);
    Update $\mathbf{Z}$ and $\theta$ by Equation (22);
    Check convergence conditions:
    $\|\mathbf{W}^{t+1} - \mathbf{W}^t\|_F^2 + \|\mathbf{E}^{t+1} - \mathbf{E}^t\|_F^2 + \|\mathbf{H}^{t+1} - \mathbf{T}^{t+1}\|_F^2 < \epsilon$;
    $t \leftarrow t + 1$;
end While
```

It can be transformedto solve each $\mathbf{T}_i$ respectively.

$$\min_{\mathbf{T}_i} \sum_{i=1}^{c} \alpha\|\mathbf{T}_i\|_* + \frac{\theta}{2}\|\mathbf{T}_i - \mathbf{H}_i + \frac{1}{\theta}\mathbf{Z}_i\|_F^2. \tag{17}$$

The optimal $\mathbf{T}_i^*$ can be obtained by the following theorem.

Theorem 2 [35] *Given* $\mathbf{Q} \in \mathbb{R}^{p\times q}$ *with rank* r and $\lambda > 0$, the optimal solution $\mathbf{A}^*$ of

$$\min_{\mathbf{A}} \lambda \|\mathbf{A}\|_* + \frac{1}{2}\|\mathbf{A} - \mathbf{Q}\|_F^2,$$

is given by $S_\lambda(\mathbf{Q})$, *and* $S_\lambda(\mathbf{Q})$ *is the following operator:*

$$\mathbf{A}^* = S_\lambda(\mathbf{Q}) = \mathbf{U}_{p\times r}\,\mathrm{diag}(\{\max(0, \delta_i(\mathbf{Q}) - \lambda)\}_{1\le i\le r})\mathbf{V}_{q\times r}^T,$$

where the singular value decomposition (SVD) of $\mathbf{Q}$ is $\mathbf{Q} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, $\boldsymbol{\Sigma} = \mathrm{diag}(\delta_1(\mathbf{Q}), \delta_2(\mathbf{Q}), \ldots, \delta_r(\mathbf{Q}))$.

According to Theorem 2, we have

$$\mathbf{T}_i^* = S_{\alpha/\theta}(\mathbf{H}_i - \mathbf{Z}_i/\theta), \tag{18}$$

where $S_{\alpha/\theta(\cdot)}$ is the soft thresholding operator. Thus we have the optimal $\mathbf{T}^* = [\mathbf{T}_1^*, \mathbf{T}_2^*, \ldots, \mathbf{T}_c^*]$.

(5) Fix others and update $\{\delta_k\}_{k=1}^v$:

$$\min_{\delta_k} \sum_{k=1}^{v} \frac{1}{\delta_k}\|\widetilde{\mathbf{W}}_k\|_F^2 \quad \text{s.t. } \delta_k > 0, \sum_{k=1}^{v}\delta_k = 1. \tag{19}$$

According to the Cauchy inequality theory [24], it holds

$$\sum_{k=1}^{v} \frac{1}{\delta_k}\|\widetilde{\mathbf{W}}_k\|_F^2 \ge \left(\sum_{k=1}^{v}\|\widetilde{\mathbf{W}}_k\|_F\right)^2, \tag{20}$$

The "=" in Equation (20) is satisfied if and only if

$$\delta_k = \frac{\|\widetilde{\mathbf{W}}_k\|_F}{\sum_{k=1}^{v}\|\widetilde{\mathbf{W}}_k\|_F}, \tag{21}$$

and Equation (19) gets minimum when δ_k takes the values of Equation (21).

(6) Fix others and update $\mathbf{Z}$, θ:

$$\begin{aligned} \mathbf{Z} &= \mathbf{Z} + \theta(\mathbf{T} - \mathbf{H}),\\ \theta &= \min(\theta_{\max}, \rho\theta), \end{aligned} \tag{22}$$

In this article, $\theta_{\max} = 10^6$ and $\rho = 1.1$. By performing steps (1)–(6) iteratively, the original loss function can be minimized until convergence or reaching the maximum iterations. Algorithm 1 summarizes the iterative algorithm for solving LWDR model in detail.

3.3 | Computational complexity analysis

From Algorithm 1, the main computational time is consumed on the calculation inside the iterations. We can observe that steps (1) and (2) contain matrix inverse operation, and step (4) contains SVD operation. The operations in other steps are addition and multiplication of matrices, which have much lower complexity than matrix inverse and SVD. The time consumption for step (1) is $O(n^3)$. In step (2), $(2 + \theta)\mathbf{I}$ is a diagonal matrix and $((2+\theta)\mathbf{I})^{-1} = \frac{1}{2+\theta}\mathbf{I}$. Then the matrix inverse operation can be simplified to matrix multiplication. Step (4) needs total c SVD operations and its time complexity is $O(c\min(ca^2, c^2a))$ where a is the average number of samples per class. Thus, the total computational complexity of Algorithm 1 is $O(t(n^3 + c\min(ca^2, c^2a)))$ if there are t iterations.

3.4 | Connections to other methods

- Connections to FLR [23]: Similar to our proposed LWDR, FLR also adopts regression model for multi-view feature learning. FLR learns a projection matrix for each view with low-rank constraint, which is helpful to explore the low-rank structure of data. However, FLR treats all views equally and ignores the fact that different views have different roles for pattern analysis, which may degrade the classification performance. Differently, our proposed LWDR adaptively learns the weights of all views and enforces those informative views to contribute more to feature learning. Thus, the proposed method can achieve better performance than FLR, which will be demonstrated in experiments.
- Connections to ADR [24]: ADR learns a weighted multi-view regression model for multi-view feature learning, which considers the different weights of different views. To avoid a rigid regression target, ADR utilizes the relaxed label matrix for regression as presented in Equation (4), which is beneficial to enlarge the margins of samples from different classes. However, according to [34], the margins of samples from the same class may be also enlarged, and the discriminative power of projection matrix will be compromised. To address this problem, LWDR introduces the class-wise low-rank constraint, which enforces the transformed samples of the same class to have the same structure. In this way, the margins of the transformed samples from the same class will be reduced and the intra-class compactness can be improved. Therefore, the proposed method has the potential to perform better than ADR.

4 | EXPERIMENTS

In this section, we conduct the experiments on Handwritten [36], Caltech101 [37], PIE [38] and AwA [39] datasets to validate the effectiveness of proposed LWDR, compared with single view and related multi-view learning methods.

FIGURE 2 Some typical images from Handwritten, PIE and Caltech101 datasets from top to bottom (images of AwA dataset are unavailable due to the copyright problem)

4.1 | Datasets

Handwritten dataset contains total 2000 images about 10 number (i.e. 0–9) with 200 images per subject. It contains six feature views, including Pixel Averages (PIX), Fourier Coefficients (FOU), Profile Correlations (FAC), Zernike Moment (ZER), Karhunen-Loeve Coefficients (KAR) and Morphological (MOR) features.

Caltech101 dataset contains 9146 images of 101 different subjects. In our experiments, total 2386 images of 20 classes are used. These images have six feature views, including Gabor, Wavelet Moments (WM), CENTRIST, HOG, GIST and LBP features.

The whole PIE dataset consists of over 40,000 face images of 68 individuals, collected under different pose, illumination and expression conditions. Total 1360 images of 68 individuals are used in our experiments, which contains five different facial poses. The five poses are used as five different views.

AwA dataset contains 30,475 images of 50 kinds of animals. In our experiment, 4000 images with 80 images per class are used. AwA dataset contains six feature views, including Global Color Histogram (GCH), Local Self-Similarity (LSS), Pyramid HOG (PHOG), RGSIFT, SIFT and SURF features.

Figure 2 shows some samples images used in our experiments. Table 1 presents the views, feature dimensions, number of classes and samples of these datasets.

4.2 | Experimental setup

For Handwritten and Caltech101 dataset, we randomly select 10 images per subject for training and the rest for testing. For PIE, five face images per person are randomly chosen for training. For AwA, the training size of each class is 20.

We compare our proposed LWDR with single view and multi-view approaches. For single view method, each view is handled separately. LDA is performed on each view for feature learning [40]. The reduced dimension of LDA is $c-1$. For multi-view methods, all views are simply concatenated and LDA is used for feature extraction, that is, LDA (all). Also, we concatenate the top three views and then perform LDA, that is, LDA (top three). The top three views are selected based on the performance of single view method.

TABLE 1 Descriptions of four datasets used in our experiments

View	Handwritten	Caltech101	PIE	AwA
V1	PIX(240)	Gabor(48)	P5(1024)	GCH(2688)
V1	FOU(76)	WM(40)	P7(1024)	LSS(2000)
V3	FAC(216)	CENTRIST(254)	P9(1024)	PHOG(252)
V4	ZER(47)	HOG(1984)	P27(1024)	RGSIFT(2000)
V5	KAR(64)	GIST(512)	P29(1024)	SIFT(2000)
V6	MOR(6)	LBP(928)	/	SURF(2000)
#classes	10	20	68	50
#samples	2000	2386	1360	4000

Multi-view methods FLR [23] and ADR [24] are performed for comparison. The parameter settings of FLR and ADR are followed the author's suggestions. Nearest neighbour with cosine distance is used for classification. We repeat the experiments 20 times by randomly sampling data partitions and report the experimental results by the mean recognition rate with standard deviation.

4.3 | Experimental results

4.3.1 | Classification accuracy comparison

Table 2 lists the classification accuracies of single view and multi-view methods on Handwritten, Caltech101, PIE and AwA datasets. As can be clearly seen, our LDWR achieves the best performance on four datasets. SVi ($i = 1, 2, \ldots, 6$) denotes the single view method which performs LDA on the i-th view. The performance of SV varies significantly, which indicates that these views have different roles for classification. The simple concatenation of all views makes effects to improve the performance and LDA (all) is superior to all SV methods. LDA (top three) can generally produce better performance the LDA (all) except Caltech101 dataset. FLR learns a projection matrix for each view and adopts majority voting for classification, which obtain better performance than simple concatenation. ADR uses adaptive weight learning strategy and ϵ-dragging technique, and it performs better than FLR. However, our proposed LWDR still outperforms ADR on multi-view feature learning. For FLR, ADR and LWDR, the tops three views of four datasets are also tested. We can observe that by using the most informative top three views, these methods can generally better performance.

4.3.2 | Adaptive weights analysis

LWDR can automatically learn the weights of different views. The large weight means its corresponding view makes more contribution in feature learning. Table 3 lists the adaptive weights learned by LWDR on all views of Handwritten, Caltech101, PIE and AwA datasets. V1, V4, V3 and V6 have the

TABLE 2 Classification accuracies of single view and multi-view feature learning methods on Handwritten, Caltech101, PIE and AwA datasets with nearest neighbour classifier. The best results are in bold

Methods	Handwritten	Caltech101	PIE	AwA
SV1	0.8247 ± 0.0296	0.5259 ± 0.0461	0.8840 ± 0.0162	0.0429 ± 0.0033
SV2	0.4082 ± 0.0459	0.4468 ± 0.0462	0.8875 ± 0.0127	0.0522 ± 0.0039
SV3	0.8749 ± 0.0242	0.3345 ± 0.0342	0.8820 ± 0.0138	0.0387 ± 0.0038
SV4	0.6013 ± 0.0407	0.7974 ± 0.0226	0.8745 ± 0.0121	0.0747 ± 0.0049
SV5	0.7341 ± 0.0362	0.7630 ± 0.0183	0.8631 ± 0.0141	0.0429 ± 0.0046
SV6	0.6199 ± 0.0184	0.8149 ± 0.0186	/	0.0799 ± 0.0042
LDA (all)	0.9122 ± 0.0087	0.9016 ± 0.0069	0.9188 ± 0.0077	0.1459 ± 0.0066
LDA (top 3)	0.9130 ± 0.0147	0.8979 ± 0.0177	0.9217 ± 0.0107	0.1493 ± 0.0053
FLR	0.9344 ± 0.0086	0.9188 ± 0.0096	0.9203 ± 0.0089	0.1607 ± 0.0073
FLR (top 3)	0.9395 ± 0.0065	0.9223 ± 0.0106	0.9314 ± 0.0145	0.1647 ± 0.0082
ADR	0.9393 ± 0.0070	0.9242 ± 0.0077	0.9258 ± 0.0086	0.1733 ± 0.0037
ADR (top 3)	0.9442 ± 0.0055	0.9398 ± 0.0162	0.9343 ± 0.0097	0.1809 ± 0.0065
LWDR	0.9415 ± 0.0088	0.9479 ± 0.0068	0.9335 ± 0.0054	0.1797 ± 0.0089
LWDR (top 3)	**0.9517** ± 0.0052	**0.9498** ± 0.0061	**0.9402** ± 0.0074	**0.1833** ± 0.0063

largest weight for the four datasets, respectively. From Table 2, these views also have good performance among all SV methods. These experimental results demonstrate that our LWDR can automatically pay more attention to those views which contain more discriminative information and make full use of them to learn discriminative features.

4.3.3 | Ablation study

In our proposed LWDR model (Equation 5), each view is assigned with an adaptively learned weight and a low-rank constraint is imposed on class-wise multi-view features to improve the intra-class compactness. To evaluate the effect of them separately, we conduct ablation experiments in this section. Two variations are derived from LWDR, that is, LWDR-s and LWDR-t. LWDR-s discards the weights learning in LWDR, that is, $\delta_k = 1$. LWDR-t discards the low-rank constraint and directly uses the label matrix to guide the feature learning. The experimental results of LWDR and its two variations on four datasets are reported in Table 4. We can observe that LWDR outperforms LWDR-s and LWDR-t, which demonstrates that the adaptive weights and class-wise low-rank constraint are effective to boost the performance.

TABLE 3 Adaptive weights learned by LWDR on four datasets

View	Handwritten	Caltech101	PIE	AwA
V1	**0.3516**	0.0324	0.1946	0.2140
V1	0.2207	0.0476	0.1990	0.2014
V3	0.0843	0.0707	**0.2118**	0.0612
V4	0.1007	**0.4613**	0.1965	0.1896
V5	0.2378	0.2085	0.1981	0.1075
V6	0.0050	0.1795	/	**0.2263**

4.3.4 | Convergence analysis

In Section 3.2, we propose an iterative algorithm for solving our LWDR model. Here we illustrate its good convergence property by experiments. Figure 4 shows the convergence curves of proposed algorithm versus iterations on Handwritten, Caltech101, PIE and AwA datasets. It is obvious that the objective function value gradually decreases to a stable value with the increase of iterations. In particular, the proposed algorithm can generally converge within 20 iterations. The experimental results demonstrate that our algorithm is effective and efficient for solving LWDR model.

4.3.5 | Parameter analysis

There are three parameters, that is, α, β, γ, in LWDR which influence the performance of proposed method. To analyse the parameter sensitivity, we first define a candidate set $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}\}$ for α and γ, and $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$ for parameter β. Then LWDR is conducted on four datasets with different combinations of the three parameters.

TABLE 4 Classification accuracies of LWDR and two variations

Method	Handwritten	Caltech101	PIE	AwA
LWDR-s	0.9285	0.9312	0.9216	0.1543
LWDR-t	0.9312	0.9387	0.9294	0.1704
LWDR	**0.9415**	**0.9479**	**0.9335**	**0.1797**

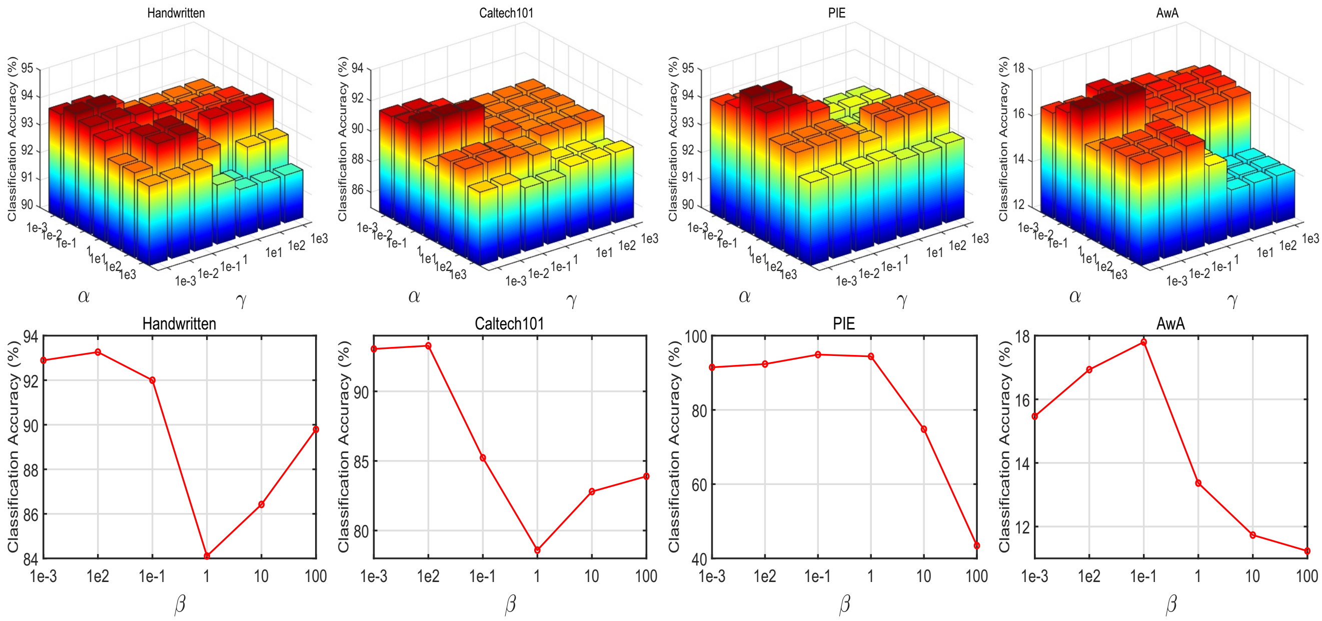

FIGURE 3 Classification accuracy (%) of LWDR versus α, β and γ on Handwritten, Caltech101, PIE and AwA datasets

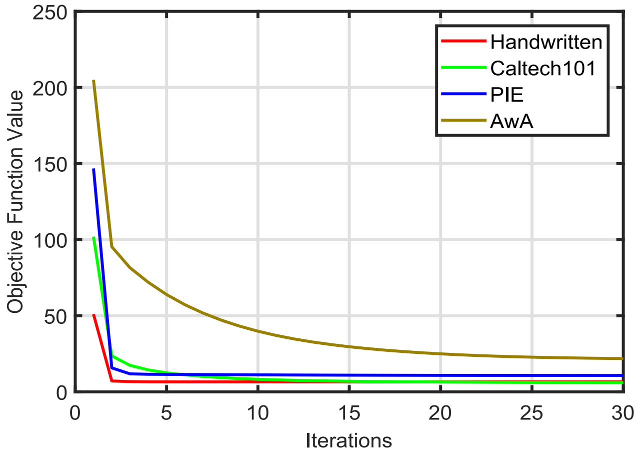

FIGURE 4 The convergence curves of proposed algorithm for solving LWDR on Handwritten, Caltech101, PIE and AwA datasets

Figure 3 shows the experimental results of LDWR versus α, γ and β on Handwritten, Caltech101, PIE and AwA datasets. It can be observed that, although all three parameters impact the performance of LWDR, LWDR can generally achieve satisfactory classification accuracy when α, γ and β locate in the range of $[10^{-2}, 10^{-1}]$, $[10^{-3}, 10^{-2}]$ and $[10^{-2}, 1]$ respectively. For example, when α, β, γ are selected from $[10^{-2}, 10^{-1}]$, $[10^{-3}, 10^{-2}]$ and $[10^{-3}, 10^{-1}]$ respectively, the classification accuracy of LWDR is satisfactory.

However, it is still difficult for optimal parameter selection on different datasets. In this article, we use the simple grid search for parameter selection [34]. We first fix β as a value in $[10^{-3}, 10^{-1}]$ like 0.01, then find the optimal α and β by different combinations of the two parameters. After obtaining the optimal α and β, we can find the optimal β by searching in its candidate set.

5 | CONCLUSION

We propose a low-rank constrained weighted discriminative regression method for multi-view feature learning (LWDR). LWDR learns a common space across all views from a unified perspective. Each view is assigned with an adaptive weight, which enables the model to focus on the important views automatically. A low-rank constraint is imposed on the features of each class, which improves the intra-class similarity and enhances the model robustness. The strict sparse label matrix is relaxed by an $l_{2,1}$-norm based regularization term. It is more flexible to deal with the errors in learning process. Experimental results on several popular datasets demonstrate the effectiveness of proposed method compared with some other single view and multi-view learning methods.

ACKNOWLEDGEMENT

The authors would like to thank the editor and anonymous reviewers for their constructive and valuable comments and suggestions. This work was partially supported by National Key Research and Development Program of China (No. 2018YFB1402600, 2016YFD0702100), and National Nature Science Foundation of China (Nos. 71671086, 71732003).

REFERENCES

1. Yang, L., et al.: Adaptive sample-level graph combination for partial multiview clustering. IEEE Trans. Image Process. 29, 2780–2794 (2020)
2. Zhu, P., et al.: Multi-view label embedding. Pattern. Recognit. 84, 126–135 (2018)
3. Guan, X., et al.: A multi-view ova model based on decision tree for multi-classification tasks. Knowl-Based. Syst. 138, 208–219 (2017)

4. Wang, Y., et al.: Multiview spectral clustering via structured low-rank matrix factorization. IEE Trans. Neural. Netw. Learn. Syst. 29(10), 4833–4843 (2018)
5. Wang, W., et al.: On deep multi-view representation learning. ICML, 1083–1092 (2015)
6. Cai, Y., et al.: Partial multi-view spectral clustering. Neurocomputing. 311, 316–324 (2018)
7. Wen, J., et al.: Generalized incomplete multi-view clustering with flexible locality structure diffusion. IEEE Trans. Cybern. (2020). https://doi.org/10.1109/TCYB.2020.2987164
8. Zhang, Z., et al.: Joint low-rank and sparse principal feature coding for enhanced robust representation and visual classification. IEEE Trans. Image Process. 25(6), 2429–2443 (2016)
9. Xu, C., Tao, D., Xu, C.: A survey on multi-view learning. arXiv preprint (2013) arXiv:13045634
10. Tao, H., et al.: Scalable multi-view semi-supervised classification via adaptive regression. IEEE Trans. Image Process. 26(9), 4283–4296 (2017)
11. Kan, M., et al.: Multi-view discriminant analysis. IEEE Trans. Pattern Anal. Mach. Intell. 38(1), 188–194 (2015)
12. Sun, S., Xie, X., Yang, M.: Multiview uncorrelated discriminant analysis. IEEE Trans. Cybern. 46(12), 3272–3284 (2016)
13. Kumar, A., Daume, H., III: A co-training approach for multi-view spectral clustering. ICML. 393–400 (2011)
14. Yu, S., et al.: Bayesian co-training. J. Mach. Learn. Res. 12, 2649–2680 (2011)
15. Kumar, A., Rai, P., Daume, H., III: Co-regularized multi-view spectral clustering. NeurIPS. 1413–1421 (2011)
16. Jiang, Y., et al.: Co-regularized plsa for multi-view clustering. ACCV. 7725, 202–213 (2012)
17. Luo, Y., et al.: Tensor canonical correlation analysis for multi-view dimension reduction. IEEE Trans. Knowl. Data Eng. 27(11), 3111–3124 (2015)
18. Zhang, Z., et al.: Robust subspace discovery by block-diagonal adaptive locality-constrained representation. ACM Multimedia, 1569–1577 (2019)
19. Liu, G., et al.: Robust subspace clustering with compressed data. IEEE Trans. Image Process. 28(10), 5161–5170 (2019)
20. Li, X., et al.: RMoR-Aion: obust multioutput regression by simultaneously alleviating input and output noises. IEEE Trans. Neural Netw. Learn. Syst. (2020). https://doi.org/10.1109/TNNLS.2020.2984635
21. Li, H., et al.: Cost-sensitive dual-bidirectional linear discriminant analysis. Inf. Sci. 510, 283–303 (2020)
22. Zhang, C., et al.: Nonnegative representation based discriminant projection for face recognition. Int. J. Mach. Learn. Cybern. (2020). https://doi.org/10.1007/s13042-020-01199-z
23. Zheng, S., et al.: A closed form solution to multi-view low-rank regression. AAAI. 1973–1979 (2015)
24. Yang, M., Deng, C., Nie, F.: Adaptive-weighting discriminative regression for multi-view classification. Pattern Recognit. 88, 236–245 (2019)
25. Hu, R., et al.: Low-rank feature selection for multi-view regression. Multimedia Tools Appl. 76(16), 17479–17495 (2017)
26. Xiang, S., et al.: Discriminative least squares regression for multiclass classification and feature selection. IEEE Trans.Neural Netw. Learn. Syst. 23(11), 1738–1754 (2012)
27. Wang, L., Pan, C.: Groupwise retargeted least-squares regression. IEEE Trans. Neural Netw. Learn. Syst. 29(4), 1352–1358 (2018)
28. Liu, G., et al.: Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern. Anal. Mach. Intell. 35(1), 171–184 (2013)
29. Zhang, Z., et al.: Deep latent low-rank fusion network for progressive subspace discovery. IJCAI, 2762–2768 (2020)
30. Li, X., et al.: Multilayer collaborative low-rank coding network for robust deep subspace discovery. ECAI. 325, 1285–1292 (2020)
31. Boyd, S., et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach. Learn. 3(1), 1–122 (2011)
32. Chen, C., et al.: The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. Math. Program. 155(1–2), 57–79 (2016)
33. Chen, C., et al.: Inertial proximal ADMM for linearly constrained separable convex optimization. SIAM J. Imaging Sci. 8(4), 2239–2267 (2015)
34. Wen, J., et al.: Inter-class sparsity based discriminative least square regression. Neural. Network. 102, 36–47 (2018)
35. Yang, J., et al.: Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. IEEE Trans. Pattern Anal. Mach. Intell. 39(1), 156–171 (2016)
36. van Breukelen, M., et al.: Handwritten digit recognition by combined classifiers. Kybernetika. 34(4), 381–386 (1998)
37. Fei, F., et al.: Learning generative visual models from few training examples: n incremental bayesian approach tested on 101 object categories. CVPR Workshop, 178–178 (2004)
38. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression database. IEEE Trans. Pattern Anal. Mach. Intell. 25(12), 1615–1618 (2003)
39. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. CVPR, 951–958 (2009)
40. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons (2012)

How to cite this article: Zhang, C., Li, H.: Low-rank constrained weighted discriminative regression for multi-view feature learning. CAAI Trans. Intell. Technol. 6(4), 471–479 (2021). https://doi.org/10.1049/cit2.12018