IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 4, 2018, accepted July 2, 2018, date of publication July 9, 2018, date of current version July 30, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2854232

Joint Linear Regression and Nonnegative Matrix Factorization Based on Self-Organized Graph for Image Clustering and Classification

WENJIE ZHU AND YUNHUI YAN

School of Mechanical Engineering and Automation, Northeastern University, Shenyang 110819, China

Corresponding author: Wenjie Zhu (wenjie_zh@126.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0304200, in part by the National Natural Science Foundation under Grant 51374063, and in part by the Fundamental Research Funds for the Central Universities under Grant N150308001.

ABSTRACT Nonnegative matrix factorization (NMF) technique has been developed successfully to represent the intuitively meaningful feature of data. A suitable representation can faithfully preserve the intrinsic structure of data. Due to the fact that it introduces the label information, semi-supervised NMF has been demonstrated more advantageous in image representation than original NMF. However, previous semi-supervised NMF variants construct a label indicator matrix only for tagging the labeled data and not being optimized together with the matrix factorization. It is short of label propagation and fails to work for predicting the attribution of data. Moreover, the transductive semi-supervised NMF variants cannot dispose the prediction of unseen data, restricting the application of NMF. In this paper, a joint optimization framework of linear regression and NMF (LR-NMF) based on the self-organized graph is proposed for a completed task which simultaneously takes into account image representation and attribution prediction. By minimizing the proposed objective, three interactive threads are running: decomposing the data into nonnegative basis matrix and the corresponding representation, linear regression using the nonnegative representation, and label propagation based on the self-organized graph which is defined in the feature space. The products of LR-NMF can be viewed as extracting nonnegative feature for clustering, meanwhile, they can be used to solve the out-of-sample problem for classification. Extensive clustering and classification experiments on the digit, face, and object challenging data sets are presented to show the efficacy of the proposed LR-NMF algorithm.

INDEX TERMS Nonnegative matrix factorization, linear regression, self-organized graph, semi-supervised clustering and classification.

I. INTRODUCTION

Matrix factorization technique has been widely investigated in image processing and pattern analysis. In the family of matrix factorization techniques, nonnegative matrix factorization (NMF) [1] has attracted more attention than others due to its ability of representing the nonnegative data such as text, image, and audio with intuitive meaning. With the effort of the experts, NMF has been used successfully in the applications of feature extraction [2]–[6], data unmixing [7], [8], image recognition [9], [10], and document clustering [11], [12]. Given the dataset $\mathbf{X} \in \mathbb{R}^{d\times n}$, NMF expects to decompose nonnegative $\mathbf{X}$ into two low-rank matrices $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{S} \in \mathbb{R}^{r\times n}$ whose elements are all nonnegative as well. $\mathbf{A}$ is called the basis matrix, and $\mathbf{S}$ is called the coefficient matrix or representation[1] of the original data over the basis matrix.

It's well known that original NMF is essentially an unsupervised method and has developed numerous variants for different application scenarios. Projective NMF which has been proposed in [13] devotes to replacing the coefficient matrix with the projective data employing the basis matrix. As a result, PNMF successfully reduces the number of parameters compared with original NMF. Then, orthogonal

[1]In this work, we treat the coefficient matrix as nonnegative feature. Hence, we treat it as nonnegative representation.

NMF (ONMF) [14], [15] introduces the orthogonal constraints into the decomposition procedure of NMF to strength the ability of sparse and parts-based representation. Another representative unsupervised NMF considers the geometric structure of the data in the procedure of NMF and this graph regularized NMF (GNMF) performs better than the mentioned unsupervised NMFs for image representation in [16]. Recently, the robust graph regularized NMF [17] is proposed for data representation to mitigate the drawback of NMF's sensitivity to outliers and noise in the data. For handling the large-scale data, the In the following decades, numerous applications based on the mentioned NMFs have been proposed for feature extraction and image clustering.

By incorporating the label information into the procedure of NMF variants, discriminative NMF methods perform well in the different classification applications. There have been numerous researches focusing on the discriminative constraint of formulation [18]–[23]. Even though, it is unlikely to encounter data with many labels in the real world. While facing to a large amount of data without labels, a subconscious way to improve the performance is labeling the data possibly. Therefore, less labeled data mixed with more unlabeled data is the real reflection on the data handed for pattern analysis task. Thus, semi-supervised learning aims at finishing the task with partial labeled data and has been getting increased attention in recent years. Lee *et al.* [24] extend the model of original NMF and propose a semi-supervised NMF (SSNMF) algorithm which employs a label indicator matrix with partial information. Subsequently, constrained NMF (CNMF) which is proposed in [25] introduces an auxiliary matrix which assists in representing the coefficient matrix with the label constraint matrix. According to CNMF, the original data matrix is decomposed into three matrices, and one of which is the label indicator matrix. Thus, CNMF converts to optimizing the basis matrix and the introduced auxiliary matrix. A label propagation based semi-supervised nonnegative matrix factorization (LpNMF) is proposed by Yi *et al.* [26]. LpNMF incorporates the label propagation into the objective of NMF in nonnegative fashion and achieves promising results on both clustering and classification. However, the predicted parameters are nonnegative and the graph is computed in the original space. Recently, another semi-supervised NMF via constraint propagation (CPSNMF) is proposed in [27] based on the framework of GNMF. Different from SSNMF, semi-supervised GNMF, or CNMF, the label indicator matrix used in [27] is not constructed with binary elements but an optimized one employing label propagation based on the graph of data before matrix factorization. Then, the multiplicative update rules of CPSNMF are given based on the framework of GNMF.

It's worth noting that all the NMF variants mentioned except LpNMF are transductive learning algorithms which fail to work for predicting unlabeled data, let alone the unseen data which are not participating in training. All of the mentioned semi-supervised NMFs except LpNMF conduct an independent process of label indicator matrix construction even that CPSNMF involves label propagation into this preprocessing. LpNMF introduces the label propagation into the optimization, however, the graph is still defined in the original space.

As for classification methods, there are a variety of approaches, such as dictionary based classification [28]–[31], dimension reduction based feature extraction [32]–[38], and promising deep learning technology [39]–[41]. Due to low demand for learning cost as well as memory space of the implementation, dimension reduction based semi-supervised learning has developed rapidly in recent years. As one of the representative algorithms, Gaussian field and harmonic function (GFHF) [32] is employed as conventional tool for semi-supervised learning with limited labeled data by introducing the label propagation technique. However, the transductive semi-supervised learning can only predict labels for unlabeled data appearing in the training data. Subsequently, semi-supervised discriminant analysis which takes account into the out-of-sample problem is proposed for semi-supervised learning. In [33], label propagation is employed into orthogonal discriminant analysis. Another dimension reduction based semi-supervised learning method in [42] constructs the virtual labels with a well-designed random walks to handle the out-of-sample problem.

Although the mentioned methods are proven to work, there are three major issues for semi-supervised learning based on the discussion of the mentioned methods:

1) The algorithms based on semi-supervised NMFs for clustering relay on the construction of label indicator matrix. No matter what strategy used, the label indicator matrix construction and NMF procedure are two independent processes, leading to suboptimal solution to the problem.
2) Semi-supervised NMFs haven't updated the label indicator matrix and fail to work for predicting the unlabeled data belonging to the training set or the unseen data never appearing in the training set.
3) The semi-supervised learning algorithms for classification aforementioned are all conducted in the space of original data which is less robust than that in the low-dimensional feature space.
4) LpNMF takes into account label propagation, however, the graph is computed in the original space which is not robust than that in the feature space. Furthermore, the label prediction is accomplished only by learning a nonnegative projection for LpNMF which is limited to the data model.

To address these issues, this work incorporates linear regression and self-organized graph regularized label propagation into the procedure of NMF in semi-supervised fashion. During learning the two nonnegative matrices, this work focuses on the following points: learning the linear regression parameters of the nonnegative representation, and label propagation based on the self-organized graph defined in the feature space to spread the label information from labeled data to unlabeled data.

TABLE 1. Notations used in this paper.

Notation	Description
$\mathbf{X}_{m\times n}$	Matrix with size of $m \times n$
$\mathbf{X}_{ij}$	The (i, j)-th element of matrix
$\mathbf{X}^{-1}$	Inverse matrix
$Tr(\mathbf{X})$	Trace of matrix
$\mathbf{X}^T$	Transformed matrix
$\|\|\mathbf{X}\|\|_F$	The Frobenius norm of matrix
$\|\mathbf{X}\|$	Element wise absolute value of matrix
$\mathbf{X}_+$	Element wise nonnegative parts of matrix
$\mathbf{X}_-$	Element wise negative parts of matrix
$\mathbf{1}$	Matrix whose elements are all one

According to integrating the constraints into the objective of this work, the proposed LR-NMF can be qualified for the following tasks:

1) Nonnegative feature extraction. With few of the amount of data labeled, the self-organized graph regularized label propagation benefits extracting more discriminative nonnegative feature than that of unsupervised NMF variants.
2) Label prediction using linear regression. LR-NMF simultaneously learns the weight and bias parameters acting on the nonnegative representation. It is robust to use the nonnegative representation rather than the original data to conduct the linear regression procedure.
3) Solving the out-of-sample problem. Semi-supervised classification confronts the problem of predicting the labels of unseen data which are not included in the training set. LR-NMF can employ the parameters of linear regression acting on the nonnegative representation of testing data to predict the label.

The remainder of this work is organized as follows: Section II introduces the related work on clustering and classification in semi-supervised fashion. The proposed LR-NMF algorithm is specified in Sec. III. Experiment results on semi-supervised clustering and classification are displayed in Sec. IV, and Section V concludes this paper.

II. RELATED WORK

Since this work is related to NMF and devotes to semi-supervised image representation, this section mainly discusses NMF variants and label propagation for semi-supervised learning. Before further introducing the related work, we specify the notations of the matrix used in this paper which are shown in Table 1.

A. NMF VARIANTS

Given a dataset matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, each column of which denotes one data point. As aforementioned, NMF aims at searching for two low-rank matrices to approximate $\mathbf{X}$:

$$\min_{\mathbf{A},\mathbf{S}} ||\mathbf{X} - \mathbf{AS}||_F^2, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{d\times r} \geqslant \mathbf{0}$ and $\mathbf{S} \in \mathbb{R}^{r\times n} \geqslant \mathbf{0}$ are the basis matrix and representation matrix, respectively. In general, $r < \min(d, n)$. The corresponding iterative multiplicative updating rules given by Lee and Seung [43] are as follows:

$$\begin{aligned}\mathbf{A}_{ik} &= \mathbf{A}_{ik}\frac{(\mathbf{XS}^T)_{ik}}{(\mathbf{ASS}^T)_{ik}},\\ \mathbf{S}_{kj} &= \mathbf{S}_{kj}\frac{(\mathbf{A}^T\mathbf{X})_{kj}}{(\mathbf{A}^T\mathbf{AS})_{kj}}.\end{aligned} \tag{2}$$

Another objective of NMF is measured using the divergence between $\mathbf{X}$ and $\mathbf{AS}$ which is given as follows:

$$\min_{\mathbf{A},\mathbf{S}\geqslant\mathbf{0}} \sum_{i,j}\left(\mathbf{X}_{ij}\log\frac{\mathbf{X}_{ij}}{(\mathbf{AS})_{ij}} - \mathbf{X}_{ij} + (\mathbf{AS})_{ij}\right), \tag{3}$$

and the corresponding multiplicative update rules are given as follows:

$$\begin{aligned}\mathbf{A}_{ik} &= \mathbf{A}_{ik}\sum_{j=1}^{n}\frac{\mathbf{X}_{ij}}{(\mathbf{AS})_{ij}}\mathbf{S}_{kj},\\ \mathbf{S}_{kj} &= \mathbf{S}_{kj}\sum_{i=1}^{m}\mathbf{A}_{ik}\frac{\mathbf{X}_{ij}}{(\mathbf{AS})_{ij}}.\end{aligned} \tag{4}$$

Among the variants of NMF, graph regularized NMF proposed by Cai *et al*. has fairly good performance on image representation. GNMF takes into account the manifold regularization during the procedure of NMF. The original optimization model of GNMF is given as:

$$\min_{\mathbf{A},\mathbf{S}\geqslant\mathbf{0}} ||\mathbf{X} - \mathbf{AS}||_F^2 + \lambda Tr(\mathbf{SL}_\mathbf{X}\mathbf{S}^T), \tag{5}$$

where $\mathbf{L}_\mathbf{X}$ is the Laplacian matrix of $\mathbf{X}$, where $\mathbf{L}_\mathbf{X} = \mathbf{D}_\mathbf{X} - \mathbf{G}_\mathbf{X}$, $(\mathbf{D}_\mathbf{X})_{ii} = \sum_j(\mathbf{G}_\mathbf{X})_{ij}$. $\mathbf{G}_\mathbf{X}$ is computed using k-nearest-neighbors technology. The trade-off parameter λ keeps a balance between the construction term and the graph preservation term. If $\lambda = 0$, the optimization degenerates to original NMF. The corresponding update rules of GNMF are given as follows:

$$\begin{aligned}\mathbf{A}_{ik} &= \mathbf{A}_{ik}\frac{(\mathbf{XS}^T)_{ik}}{(\mathbf{ASS}^T)_{ik}},\\ \mathbf{S}_{kj} &= \mathbf{S}_{kj}\frac{(\mathbf{A}^T\mathbf{X} + \lambda\mathbf{SG}_\mathbf{X})_{kj}}{(\mathbf{A}^T\mathbf{AS} + \lambda\mathbf{SD}_\mathbf{X})_{kj}}.\end{aligned} \tag{6}$$

The unsupervised NMFs including PNMF, ONMF, and GNMF cannot assist in obtaining discriminative image representation. For the sake of employing the label information, semi-supervised NMF incorporates the label information into the matrix factorization procedure. One of the representative semi-supervised NMF, CNMF introduces a label constraint matrix and an auxiliary matrix, aiming at decomposing the original matrix into three parts:

$$\min_{\mathbf{A},\mathbf{Z}\geqslant\mathbf{0}} ||\mathbf{X} - \mathbf{AZT}||_F^2, \tag{7}$$

where $\mathbf{ZT} \approx \mathbf{S}$, $\mathbf{Z} \in \mathbb{R}^{r\times(c+n-l)}$ is the introduced auxiliary matrix, and $\mathbf{T} \in \mathbb{R}^{(c+n-l)\times n}$ is the label indicator matrix,

where c is the number of classes, and l is the number of labeled data. $\mathbf{T}$ is a binary matrix, e.g. $\mathbf{T}_{ij} = 1$ if the j-th data belongs to the i-th class, and the unlabeled partition of $\mathbf{T}$ is padded with an identity matrix with size of $(n-l)\times(n-l)$. Then, the multiplicative updating rules of $\mathbf{A}$ and $\mathbf{Z}$ are given in [25] in detail.

B. LABEL PROPAGATION

Label propagation (LP) is a well-known semi-supervised learning method [32]. According to the theory of LP, the nearby data points or the data points belonging to the same cluster should share same labels. Given the original data matrix $\mathbf{X} = \{\mathbf{X}_l, \mathbf{X}_u\}$, where $\mathbf{X}_l \in \mathbb{R}^{d\times l}$ denotes the labeled data and $\mathbf{X}_u \in \mathbb{R}^{d\times u}$ denotes the unlabeled one. The corresponding label prediction matrix is denoted as $\mathbf{F}_l \in \mathbb{R}^{l\times c}$ and $\mathbf{F}_u \in \mathbb{R}^{u\times c}$, respectively. Thus, we have $n = l + u$ and the label prediction matrix of total data points is denoted as $\mathbf{F} = \begin{bmatrix}\mathbf{F}_l \\ \mathbf{F}_u\end{bmatrix}$. Moreover, the graph of the original data is adopted in the optimization of LP. The graph $\mathbf{G}_\mathbf{X} \in \mathbb{R}^{n\times n}$ is calculated to encode the similarities between data pairs. In addition, the label indicator matrix of the partial data is given as $\mathbf{Y}_l \in \mathbb{R}^{l\times c}$. In practice, the labeled data is insufficient but reliable. Thus, it's almost certain that $\mathbf{F}_l = \mathbf{Y}_l$. The optimization of LP can be defined as follows:

$$\min_{\mathbf{F}_u} \frac{1}{2}\sum_{i,j}^{n}(\mathbf{G}_\mathbf{X})_{ij}||\mathbf{F}_i - \mathbf{F}_j||_F^2, \quad \text{s.t.} \quad \mathbf{F}_l = \mathbf{Y}_l, \tag{8}$$

where $\mathbf{G}$ is same with that of GNMF. Thus, we can employ the same strategy and reformulate the objective of LP which is shown as follows:

$$\min_{\mathbf{F}_u} Tr(\mathbf{F}^T\mathbf{L}_\mathbf{X}\mathbf{F}) \quad \text{s.t. } \mathbf{F}_l = \mathbf{Y}_l. \tag{9}$$

III. PROPOSED METHOD

As discussed before, this paper devotes to generating a fully functional NMF which can be qualified for image clustering and classification with few labeled data. The original NMF only obtains the nonnegative representation in unsupervised fashion. Semi-supervised NMFs put emphasis on constructing the label indicator matrix which is independent from the procedure of NMF. Moreover, it is short of predicting the labels of data, even solving the out-of-sample problem for classification. Even though there have been numerous algorithms [33], [42], [44] focusing on semi-supervised learning which can also work for unseen data, they conduct the learning procedure in the original data space rather than low-dimensional and optimal feature space. To address the issues, the proposed optimization framework of NMF aims at jointly learning the nonnegative representation and the parameters of linear regression for semi-supervised image clustering and classification. In this section, we first propose the objective of LR-NMF, and then give the optimization strategy, complexity and convergence analysis, and the procedure of semi-supervised learning employing the algorithm of LR-NMF.

A. OBJECTIVE OF LR-NMF

Similar to the definition of LP, let $\mathbf{X} \in \mathbb{R}^{d\times n}$ denote the original data matrix which consists of labeled data $\mathbf{X}_l \in \mathbb{R}^{d\times l}$ and unlabeled data $\mathbf{X}_u \in \mathbb{R}^{d\times u}$, where d is the dimension of data and $l < u$ in general. Denote the label prediction matrix by $\mathbf{F} \in \{0,1\}^{n\times c}$. In order to take advantage of the label information, we first formulate the objective of semi-supervised NMF which introduces the label propagation constraint. It is expected to explore the latent discriminative information using the relationships of the feature data points. Thus, the optimization problem involved label propagation acting on the nonnegative representation is given as follows:

$$\begin{aligned}\min_{\mathbf{A},\mathbf{S},\mathbf{F}_u\geqslant 0} &\left\{\begin{matrix}||\mathbf{X}-\mathbf{A}\mathbf{S}||_F^2 + \alpha Tr(\mathbf{F}^T\mathbf{L}\mathbf{F}) \\ +\lambda||\mathbf{A}^T\mathbf{X} - \mathbf{A}^T\mathbf{X}\mathbf{G}||_F^2\end{matrix}\right\}, \\ &\text{s.t. } \mathbf{F}_l = \mathbf{Y}_l, \quad \mathbf{G}\geqslant\mathbf{0},\ \mathbf{G}_{ii} = 0,\ ||\mathbf{G}_i||_0 \leqslant K,\end{aligned} \tag{10}$$

In (10), α and λ are the regularized parameters. It has to be emphasized that $\mathbf{L}$ is different from that defined in the objective of GNMF or LP. The Laplacian matrix $\mathbf{L}$ is computed via the sparse representation $\mathbf{G}$ which can be achieved by optimizing the second third term of the objective in (10) using the projected, nonnegative feature set $\{\mathbf{A}^T\mathbf{X}_j\}$. As emphasized in the proposed method, the semi-supervised learning is conducted in the low-dimensional feature space. Therefore, $\mathbf{L} = \mathbf{D} - \widehat{\mathbf{G}}$, $\mathbf{D}_{ii} = \sum_j(\widehat{\mathbf{G}})_{ij}$, and $\widehat{\mathbf{G}} = (\mathbf{G}+\mathbf{G}^T)/2$. Since the graph $\widehat{\mathbf{G}}$ is self-organized, the nonnegative, sparse, and diagonal constraints are involved into the optimization in (10). Furthermore, K is the sparsity and $||\mathbf{G}_i||_0 \leqslant K$ indicates there are K non-zero elements at most for each column of $\mathbf{G}$.

Although (10) is competent to semi-supervised NMF for feature extraction, the regularization of LP is an transductive way and could only predict the labels of training data. Thus, it is necessary to learn the parameters of predicting the labels of unseen data which have not participated at the training phase. For this, a linear regression acting on the nonnegative representation is conducted during the optimization. We define the objective of this work which is shown in (11).

$$\begin{aligned}\min_{\mathbf{A},\mathbf{S},\mathbf{W},\mathbf{b},\mathbf{F}_u} &\left\{\begin{matrix}||\mathbf{X}-\mathbf{A}\mathbf{S}||_F^2 + \alpha Tr(\mathbf{F}^T\mathbf{L}\mathbf{F}) \\ +\lambda||\mathbf{A}^T\mathbf{X} - \mathbf{A}^T\mathbf{X}\mathbf{G}||_F^2 + \\ \beta||\mathbf{S}^T\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{F}||_F^2 + \gamma||\mathbf{W}||_F^2\end{matrix}\right\}, \\ &\text{s.t. } \mathbf{F}_l = \mathbf{Y}_l, \quad \mathbf{A},\mathbf{S},\mathbf{G}\geqslant\mathbf{0},\ \mathbf{G}_{ii} = 0,\ ||\mathbf{G}_i||_0 \leqslant K.\end{aligned} \tag{11}$$

In (11), $\mathbf{W} \in \mathbb{R}^{r\times c}$ and $\mathbf{b} \in \mathbb{R}^{c\times 1}$ are the weight and bias parameters of linear regression, $\mathbf{1} \in \mathbb{R}^{n\times 1}$, and β and γ are the regularized parameters. The linear regression aims at learning $\{\mathbf{W}, \mathbf{b}\}$ to make $\mathbf{W}^T\mathbf{S}_i + \mathbf{b} \approx \mathbf{F}_i$ hold. By doing this, it could be convenient to adopt the learned parameters to predict any data points with same distribution of the training data. In the objective of LR-NMF, the first term is the construction error between original data and the product of basis and representation. The second term employs LP based on the self-organized graph for semi-supervised learning. The third

term embeds data into a low-dimensional space and simultaneously finds the sparse code in this space. The fourth term regularizes the linear approximation error and the last one avoids from over-fitting. By integrating these constraints into the objective of LR-NMF, it is expected to exploit alternately updating algorithm to solve the optimization problem of LR-NMF for joint feature extraction and predicting labels of data.

B. OPTIMIZATION

The unknown variables in the objective of LR-NMF consists of the basis $\mathbf{A}$, representation $\mathbf{S}$, nonnegative sparse representation $\mathbf{G}$, label indicator matrix of unlabeled data $\mathbf{F}_u$, weight $\mathbf{W}$ and bias $\mathbf{b}$. Although the objective is not convex to the variables $\{\mathbf{A}, \mathbf{S}, \mathbf{W}, \mathbf{b}, \mathbf{F}_u\}$, it is solvable to update one while fixing others alternately.

1) UPDATING $\{\mathbf{A}, \mathbf{S}\}$ WHILE FIXING $\{\mathbf{G}, \mathbf{W}, \mathbf{b}, \mathbf{F}_u\}$

The optimization problem with respect to $\mathbf{A}$ and $\mathbf{S}$ is rewritten as follows:

$$\min_{\mathbf{A},\mathbf{S}} \left\{ \begin{array}{c} ||\mathbf{X}-\mathbf{AS}||_F^2 + \lambda||\mathbf{A}^T\mathbf{X}-\mathbf{A}^T\mathbf{XG}||_F^2 \\ +\beta||\mathbf{S}^T\mathbf{W}+\mathbf{1b}^T-\mathbf{F}||_F^2 \end{array} \right\}, \quad \text{s.t. } \mathbf{A} \geqslant \mathbf{0}, \quad \mathbf{S} \geqslant \mathbf{0}. \tag{12}$$

Following the similar way as [1], [25], and [43], the iterative updating algorithm is employed in this paper. The functions with respect to $\mathbf{A}$ and $\mathbf{S}$ are expressed as:

$$J_{\mathbf{A}} = ||\mathbf{X}-\mathbf{AS}||_F^2 + \lambda||\mathbf{A}^T\mathbf{X}-\mathbf{A}^T\mathbf{XG}||_F^2, \tag{13}$$

and

$$J_{\mathbf{S}} = ||\mathbf{X}-\mathbf{AS}||_F^2 + \beta||\mathbf{S}^T\mathbf{W}+\mathbf{1b}^T-\mathbf{F}||_F^2. \tag{14}$$

To further solutions, we first rewrite the objective functions in the form of matrix trace:

$$\begin{aligned} J_{\mathbf{A}} &= Tr((\mathbf{X}-\mathbf{AS})^T(\mathbf{X}-\mathbf{AS})) \\ &\quad + \lambda Tr(\mathbf{A}^T(\mathbf{X}-\mathbf{XG})(\mathbf{X}-\mathbf{XG})^T\mathbf{A}) \\ &= Tr(\mathbf{X}^T\mathbf{X}-2\mathbf{X}^T\mathbf{AS}+\mathbf{SS}^T\mathbf{A}^T\mathbf{A}) \\ &\quad + \lambda Tr((\mathbf{X}-\mathbf{XG})(\mathbf{X}-\mathbf{XG})^T\mathbf{AA}^T), \end{aligned} \tag{15}$$

and

$$\begin{aligned} J_{\mathbf{S}} &= Tr((\mathbf{X}-\mathbf{AS})^T(\mathbf{X}-\mathbf{AS})) \\ &\quad + \beta((\mathbf{S}^T\mathbf{W}+\mathbf{1b}^T-\mathbf{F})^T(\mathbf{S}^T\mathbf{W}+\mathbf{1b}^T-\mathbf{F})) \\ &= Tr(\mathbf{X}^T\mathbf{X}-2\mathbf{X}^T\mathbf{AS}+\mathbf{SS}^T\mathbf{A}^T\mathbf{A}) \\ &\quad + \beta Tr(\mathbf{SS}^T\mathbf{WW}^T+2(\mathbf{1b}^T-\mathbf{F})\mathbf{W}^T\mathbf{S}+\mathbf{F}^T\mathbf{F}). \end{aligned} \tag{16}$$

Omitting the irrelevant terms, we define the auxiliary matrix $\Phi \geqslant \mathbf{0}$ and $\Psi \geqslant \mathbf{0}$, and then give the Lagrange function with respect to $\mathbf{A}$ and $\mathbf{S}$ as:

$$\begin{aligned} \mathcal{L}_{\mathbf{A}} &= Tr(\mathbf{SS}^T\mathbf{A}^T\mathbf{A}-2\mathbf{X}^T\mathbf{AS}) \\ &\quad + \lambda Tr((\mathbf{X}-\mathbf{XG})(\mathbf{X}-\mathbf{XG})^T\mathbf{AA}^T) + Tr(\Phi\mathbf{A}^T), \end{aligned} \tag{17}$$

and

$$\begin{aligned} \mathcal{L}_{\mathbf{S}} &= Tr(\mathbf{SS}^T\mathbf{A}^T\mathbf{A}-2\mathbf{X}^T\mathbf{AS}) + \beta Tr(\mathbf{SS}^T\mathbf{WW}^T \\ &\quad + 2(\mathbf{1b}^T-\mathbf{F})\mathbf{W}^T\mathbf{S}) + Tr(\Psi\mathbf{S}^T). \end{aligned} \tag{18}$$

Then, we compute the derivative of Lagrange function to $\mathbf{A}$ and $\mathbf{S}$:

$$\frac{\partial \mathcal{L}_{\mathbf{A}}}{\partial \mathbf{A}} = -2\mathbf{XS}^T + 2\mathbf{ASS}^T + 2\lambda(\mathbf{X}-\mathbf{XG})(\mathbf{X}-\mathbf{XG})^T\mathbf{A} + \Phi. \tag{19}$$

and

$$\begin{aligned} \frac{\partial \mathcal{L}_{\mathbf{S}}}{\partial \mathbf{S}} &= -2\mathbf{A}^T\mathbf{X} + 2\mathbf{A}^T\mathbf{AS} + 2\beta\mathbf{WW}^T\mathbf{S} \\ &\quad + 2\beta\mathbf{Wb1}^T - 2\beta\mathbf{WF}^T + \Psi. \end{aligned} \tag{20}$$

Based on the Karush-Kuhn-Tucker condition that $\Phi_{ij}\mathbf{A}_{ij} = 0$ and $\Psi_{ij}\mathbf{S}_{ij} = 0$, we set the derivative to zero and have the following equation:

$$\begin{aligned} &(\mathbf{ASS}^T)_{ij}\mathbf{A}_{ij} - (\mathbf{XS}^T)_{ij}\mathbf{A}_{ij} \\ &\quad + \lambda((\mathbf{XX}^T - 2\mathbf{XGX}^T + \mathbf{XGG}^T\mathbf{X}^T)\mathbf{A})_{ij}\mathbf{A}_{ij} = 0, \end{aligned} \tag{21}$$

and

$$\begin{aligned} &(\mathbf{A}^T\mathbf{AS})_{ij}\mathbf{S}_{ij} + \beta(\mathbf{WW}^T\mathbf{S})_{ij}\mathbf{S}_{ij} + \beta(\mathbf{Wb1}^T)_{ij}\mathbf{S}_{ij} \\ &\quad - (\mathbf{A}^T\mathbf{X})_{ij}\mathbf{S}_{ij} - \beta(\mathbf{WF}^T)_{ij}\mathbf{S}_{ij} = 0. \end{aligned} \tag{22}$$

Therefore, we can get the update rule of $\mathbf{A}$:

$$\mathbf{A}_{ij} = \mathbf{A}_{ij}\frac{(\mathbf{XS}^T + 2\lambda\mathbf{XGX}^T\mathbf{A})_{ij}}{(\mathbf{ASS}^T + \lambda(\mathbf{XX}^T\mathbf{A} + \mathbf{XGG}^T\mathbf{X}^T\mathbf{A}))_{ij}}. \tag{23}$$

Let $\mathbf{W}_+$, $\mathbf{b}_+$ and $\mathbf{F}_+$ denote the nonnegative parts of matrix $\mathbf{W}$, $\mathbf{b}$ and $\mathbf{F}$. Then, $\mathbf{W}_-$, $\mathbf{b}_-$ and $\mathbf{F}_-$ denote the negative parts of corresponding matrix in order to satisfy the additive rule. Hence, we substitute them into the equation above. Then, we can get the updating rule of $\mathbf{S}$ which is presented as follows:

$$\mathbf{S}_{ij} = \mathbf{S}_{ij}\frac{(\mathbf{A}^T\mathbf{X} + \beta\mathbf{W}_+\mathbf{M} + \beta|\mathbf{W}_-|\mathbf{N})_{ij}}{(\mathbf{A}^T\mathbf{AS} + \beta|\mathbf{W}_-|\mathbf{M} + \beta\mathbf{W}_+\mathbf{N})_{ij}}, \tag{24}$$

where $\mathbf{M} = |\mathbf{W}_-|^T\mathbf{S} + |\mathbf{b}_-|\mathbf{1}^T + |\mathbf{F}_-|^T$ and $\mathbf{N} = \mathbf{W}_+^T\mathbf{S} + \mathbf{b}_+\mathbf{1}^T + \mathbf{F}_+^T$.

2) UPDATING LINEAR REGRESSION PARAMETERS $\{\mathbf{W}, \mathbf{b}\}$ WHILE FIXING $\mathbf{S}$ AND $\mathbf{F}_u$

With fixed $\mathbf{S}$ and $\mathbf{F}_u$, it falls into a regularized linear regression problem which aims at searching for the relation between the nonnegative representation and the corresponding labels. The objective to $\mathbf{b}$ is formulated as $J_{\mathbf{b}} = ||\mathbf{S}^T\mathbf{W}+\mathbf{1b}^T-\mathbf{F}||_F^2$. We first fix $\mathbf{W}$ and set the derivative w.r.t. $\mathbf{b}$ to zero, and then we can get the closed solution to $\mathbf{b}$ as follows:

$$\mathbf{b} = \frac{(\mathbf{F}^T - \mathbf{W}^T\mathbf{S})\mathbf{1}}{\mathbf{1}^T\mathbf{1}}. \tag{25}$$

We substitute (25) into the objective with respect to $\mathbf{W}$:

$$J_{\mathbf{W}} = \beta||(\mathbf{I} - \frac{\mathbf{11}^T}{\mathbf{1}^T\mathbf{1}})(\mathbf{S}^T\mathbf{W} - F)||_F^2 + \gamma||\mathbf{W}||_F^2. \tag{26}$$

Let the centering matrix $\mathbf{H} = (\mathbf{I} - \frac{\mathbf{11}^T}{\mathbf{1}^T\mathbf{1}})^T(\mathbf{I} - \frac{\mathbf{11}^T}{\mathbf{1}^T\mathbf{1}})$, we rewrite the objective in form of the trace of matrix as $J_{\mathbf{W}} = \beta Tr((\mathbf{S}^T\mathbf{W} - \mathbf{F})\mathbf{H}(\mathbf{S}^T\mathbf{W} - \mathbf{F}))$. By enforcing the derivative to zero and have the closed-form solution to $\mathbf{W}$:

$$\mathbf{W} = \beta(\beta\mathbf{SHS}^T + \gamma\mathbf{I})^{-1}\mathbf{SHF}. \tag{27}$$

3) UPDATING $\mathbf{F}_u$ WHILE FIXING $\mathbf{S}$

Before updating $\mathbf{F}_u$, we can construct the graph via (8), and then compute $\mathbf{L}$. From the solution of $\mathbf{W}$, it can be observed that $\mathbf{W}$ can be represented using $\mathbf{F}$ over the projection $\beta(\beta\mathbf{S}\mathbf{H}\mathbf{S}^T + \gamma\mathbf{I})^{-1}\mathbf{S}\mathbf{H}$. Since that $\mathbf{F}$ is related to $\mathbf{W}$, we reorganize the objective to $\mathbf{F}$ which includes the unsolved $\mathbf{F}_u$ as follows:

$$J_{\mathbf{F}} = \alpha Tr(\mathbf{F}^T\mathbf{L}\mathbf{F}) + J, \tag{28}$$

where

$$\begin{aligned} J &= \beta Tr((\mathbf{W}^T\mathbf{S} - \mathbf{F}^T)\mathbf{H}(\mathbf{S}^T\mathbf{W} - \mathbf{F})) + \gamma Tr(\mathbf{W}^T\mathbf{W}) \\ &= Tr(\beta(\mathbf{F}^T\mathbf{H}\mathbf{F} - 2\mathbf{F}^T\mathbf{H}\mathbf{S}^T\mathbf{W}) + \mathbf{W}^T(\beta\mathbf{S}\mathbf{H}\mathbf{S}^T + \gamma\mathbf{I})\mathbf{W}) \\ &= \beta Tr(\mathbf{F}^T\mathbf{H}\mathbf{F}) - 2\beta Tr(\mathbf{F}^T\mathbf{H}\mathbf{S}^T\mathbf{W}) + \beta Tr(\mathbf{W}^T\mathbf{S}\mathbf{H}\mathbf{F}) \\ &= \beta(Tr(\mathbf{F}^T\mathbf{H}\mathbf{F}) - \beta Tr(\mathbf{F}^T\mathbf{H}\mathbf{S}^T\mathbf{W})) \\ &= Tr(\mathbf{F}^T(\beta\mathbf{H} - \beta^2\mathbf{H}\mathbf{S}^T(\beta\mathbf{S}\mathbf{H}\mathbf{S}^T + \gamma\mathbf{I})^{-1}\mathbf{S}\mathbf{H})\mathbf{F}). \end{aligned} \tag{29}$$

Substituting (29) into (28), we can get the optimization problem of $\mathbf{F}$ as follows:

$$\min_{\mathbf{F}_u} Tr(\mathbf{F}^T\mathbf{R}\mathbf{F}) \quad \text{s.t. } \mathbf{F}_l = \mathbf{Y}_l, \tag{30}$$

where $\mathbf{R} = \alpha\mathbf{L} + \beta\mathbf{H} - \beta^2\mathbf{H}\mathbf{S}^T(\beta\mathbf{S}\mathbf{H}\mathbf{S}^T + \gamma\mathbf{I})^{-1}\mathbf{S}\mathbf{H}$. Then, we partition $\mathbf{R}$ into four blocks with proper sizes and substitute $\mathbf{F}_l = \mathbf{Y}_l$ into the objective of (30). The objective with respect to $\mathbf{F}_u$ is rewritten as:

$$\begin{aligned} J_{\mathbf{F}_u} &= Tr(\begin{bmatrix}\mathbf{Y}_l \\ \mathbf{F}_u\end{bmatrix}^T \begin{bmatrix}\mathbf{R}_{ll} & \mathbf{R}_{lu} \\ \mathbf{R}_{ul} & \mathbf{R}_{uu}\end{bmatrix} \begin{bmatrix}\mathbf{Y}_l \\ \mathbf{F}_u\end{bmatrix}) \\ &= Tr(\mathbf{F}_u\mathbf{R}_{ul}\mathbf{Y}_l + \mathbf{F}_u\mathbf{R}_{uu}\mathbf{F}_u), \end{aligned} \tag{31}$$

where $\mathbf{R}_{ll} \in \mathbb{R}^{l\times l}$, $\mathbf{R}_{lu} \in \mathbb{R}^{l\times u}$, $\mathbf{R}_{ul} \in \mathbb{R}^{u\times l}$, and $\mathbf{R}_{uu} \in \mathbb{R}^{u\times u}$. Then, the optimization of $\mathbf{F}_u$ is relevant for $\mathbf{R}_{ul}$, $\mathbf{R}_{uu}$, and $\mathbf{Y}_l$. The optimal solution $\mathbf{F}_u$ can be obtained by setting the derivative of $J_{\mathbf{F}_u}$ with respect to $\mathbf{F}_u$ to zero:

$$\mathbf{F}_u = -\mathbf{R}_{uu}^{-1}\mathbf{R}_{ul}\mathbf{Y}_l. \tag{32}$$

4) UPDATING $\mathbf{G}$ WHILE FIXING $\{\mathbf{F}, \mathbf{A}\}$

The solution of $\mathbf{G}$ can be obtained by solving the following problem:

$$\min J_{\mathbf{G}} = \alpha Tr(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \lambda||\mathbf{A}^T\mathbf{X} - \mathbf{A}^T\mathbf{X}\mathbf{G}||_F^2 \tag{33}$$

According to Lemma 1 which is proved in [45], $\mathbf{L}$ is related to $\mathbf{G}$.

Lemma 1: Given an affinity matrix $\widehat{\mathbf{G}} \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\mathbf{D}$ defined as $\mathbf{D}_{ii} = \sum_j \widehat{\mathbf{G}}_{ij}$, for its Laplacian matrix $\mathbf{L} \in \mathbb{R}^{n\times n}$ defined as $\mathbf{L} = \mathbf{D} - \widehat{\mathbf{G}}$ and a matrix $\mathbf{F} \in \mathbb{R}^{n\times c}$, we have

$$Tr(\mathbf{F}^T\mathbf{L}\mathbf{F}) = \frac{1}{2}Tr(\widehat{\mathbf{G}}\mathbf{Q}), \tag{34}$$

where $\mathbf{Q}_{ij} = ||\mathbf{F}^i - \mathbf{F}^j||^2$, and $\mathbf{F}^j$ is the j-th row of matrix $\mathbf{F}$. It's well-known that $Tr(X) = Tr(X^T)$, $Tr(XY) = Tr(YX)$, and $\mathbf{Q} = \mathbf{Q}^T$. We denote $\mathbf{Z} = \mathbf{A}^T\mathbf{X}$ and have the objective w.r.t. $\mathbf{G}$:

$$\begin{aligned} J_{\mathbf{G}} &= \alpha Tr(\mathbf{F}^T\mathbf{L}\mathbf{F}) + \lambda||\mathbf{A}^T\mathbf{X} - \mathbf{A}^T\mathbf{X}\mathbf{G}||_F^2 \\ &= \frac{\alpha}{2}Tr(\widehat{\mathbf{G}}\mathbf{Q}) + \lambda||\mathbf{Z} - \mathbf{Z}\mathbf{G}||_F^2 \\ &= \frac{\alpha}{4}Tr((\mathbf{G} + \mathbf{G}^T)\mathbf{Q}) + \lambda||\mathbf{Z} - \mathbf{Z}\mathbf{G}||_F^2 \\ &= \frac{\alpha}{2}Tr(\mathbf{Q}\mathbf{G}) + \lambda||\mathbf{Z} - \mathbf{Z}\mathbf{G}||_F^2. \end{aligned} \tag{35}$$

Thus, $\mathbf{G}$ can be achieved by solving the following problem:

$$\begin{aligned} &\min_{\mathbf{G}} \lambda||\mathbf{Z} - \mathbf{Z}\mathbf{G}||_F^2 + \frac{\alpha}{2}Tr(\mathbf{G}\mathbf{Q}) \\ &\text{s.t. } \mathbf{G} \geqslant \mathbf{0}, \quad \mathbf{G}_{ii} = 0,\ ||\mathbf{G}_i||_0 \leqslant K. \end{aligned} \tag{36}$$

It can be further decoupled to solve each column $\mathbf{G}_i$ separately as follows,

$$\begin{aligned} &\min_{\mathbf{G}_i} \lambda||\mathbf{Z}_i - \mathbf{Z}\mathbf{G}_i||_F^2 + \frac{\alpha}{2}\mathbf{Q}^i\mathbf{G}_i \\ &\text{s.t. } \mathbf{G}_i \geqslant 0, \quad \mathbf{G}_{ii} = 0,\ ||\mathbf{G}_i||_0 \leqslant K. \end{aligned} \tag{37}$$

The problem (39) is a weighted ℓ_0 norm optimization and can be solved by projected gradient descent method as shown in Algorithm 1.

Algorithm 1 Algorithm for Getting Each Column of $\mathbf{G}$

Input: $\mathbf{Z}$, $\mathbf{Q}^i$, λ, α, K, $\eta > 0$.
Output: $\mathbf{G}_i$
if ***not converged*** **then**
  *Step*1 : Compute the gradient :
  $\nabla_{\mathbf{G}_i} = 2\lambda\mathbf{Z}^T(\mathbf{Z}_i - \mathbf{Z}\mathbf{G}_i) + \frac{\alpha}{2}\mathbf{Q}_i$;
  *Step*2 : Gradient projection:
  $\mathbf{G}_i = \max(\mathbf{G}_i - \eta \cdot \nabla_{\mathbf{G}_i}, 0)$;
  *Step*3 : If $i \neq j$ and $\mathbf{G}_{ij}$ is the largest K value of $\mathbf{G}_i$:
  $\mathbf{G}_{ij} = \mathbf{G}_{ij}$, otherwise, $\mathbf{G}_{ij} = 0$.
end

While conducting the algorithm of LR-NMF, the variables can be updated alternatively until the value of the objective does not change. The complete algorithm of LR-NMF is outlined in Algorithm 2. For better readability, we use $\odot$ and $\oslash$ denote the element wise multiplication and division operators.

C. COMPLEXITY AND CONVERGENCE ANALYSIS

In general, the major computation cost lies in multiplication updating rules of $\mathbf{A}$ and $\mathbf{S}$. As for updating others, the solutions of $\mathbf{W}$ and $\mathbf{F}_u$ involves inverse matrix. If the algorithm of LR-NMF employs linear equations system to compute the inverse matrix when computing $\mathbf{W}$ and $\mathbf{F}_u$, the complexity will be reduced greatly based on the observation that $\beta\mathbf{S}\mathbf{H}\mathbf{S}^T + \gamma\mathbf{I}$ and $\mathbf{R}_{uu}$ are sparse, positive definite, and symmetric. Besides, the computation complexity of updating $\{\mathbf{W}, \mathbf{b}, \mathbf{F}_u\}$ could be trivial owing their closed-form solutions. The computation complexity of Algorithm 2 for solving the

Algorithm 2 Algorithm Outline for the Optimization of LR-NMF

Input: dataset matrix $\mathbf{X} = \{\mathbf{X}_l, \mathbf{X}_u\}$; Label indicator matrix $\mathbf{Y}_l$ for labeled data $\mathbf{X}_l$; Parameters α, β, λ, and γ.
Output: Bases $\mathbf{A}$; Representations $\mathbf{S}$; Predicted label indicator matrix $\mathbf{F}_u$; Weight $\mathbf{W}$; Bias $\mathbf{b}$.
Initialize $\mathbf{A}^0 \in \mathbb{R}^{m\times r} \geqslant \mathbf{0}$; $\mathbf{S}^0 \in \mathbb{R}^{r\times n} \geqslant \mathbf{0}$; $\mathbf{W}^0 \in \mathbb{R}^{c\times c}$; $\mathbf{b}^0 \in \mathbb{R}^{c\times 1}$; $\mathbf{G}^0 \in \mathbb{R}^{n\times n}$;
if *not converged* **then**
  *Step*1 : Update the basis matrix via : $\mathbf{A} = \mathbf{A} \odot ((\mathbf{X}\mathbf{S}^T + 2\lambda\mathbf{X}\mathbf{G}\mathbf{X}^T\mathbf{A}) \oslash (\mathbf{A}\mathbf{S}\mathbf{S}^T + \lambda(\mathbf{X}\mathbf{X}^T + \mathbf{X}\mathbf{G}\mathbf{G}^T\mathbf{X}^T)\mathbf{A}))$;
  *Step*2 : Compute $\mathbf{M} = |\mathbf{W}_-|^T\mathbf{S} + |\mathbf{b}_-|\mathbf{1}^T + |\mathbf{F}_-|^T$, $\mathbf{N} = \mathbf{W}_+^T\mathbf{S} + \mathbf{b}_+\mathbf{1}^T + \mathbf{F}_+^T$ and Update the nonnegative representation via : $\mathbf{S} = \mathbf{S} \odot ((\mathbf{A}^T X + \beta\mathbf{W}_+\mathbf{M} + \beta|\mathbf{W}_-|\mathbf{N}) \oslash (\mathbf{A}^T\mathbf{A}\mathbf{S} + \beta|\mathbf{W}_-|\mathbf{M} + \beta\mathbf{W}_+\mathbf{N}))$;
  *Step*3 : Construct the Laplacian matrix $\mathbf{L}$ via $\mathbf{L} = \mathbf{D} - \widehat{\mathbf{G}}$ and $\widehat{\mathbf{G}} = (\mathbf{G} + \mathbf{G}^T)/2$;
  *Step*4 : Compute the temporary matrix : $\mathbf{R} = \alpha\mathbf{L} + \beta\mathbf{H} - \beta^2\mathbf{H}\mathbf{S}^T(\beta\mathbf{S}\mathbf{H}\mathbf{S}^T + \gamma\mathbf{I})^{-1}\mathbf{S}\mathbf{H}$;
  *Step*5 : Update label indicator matrix of unlabeled data via : $\mathbf{F}_u = -\mathbf{R}_{uu}^{-1}\mathbf{R}_{ul}\mathbf{Y}_l$, and construct $\mathbf{F} = \begin{bmatrix}\mathbf{Y}_l \\ \mathbf{F}_u\end{bmatrix}$;
  *Step*6 : Update $\mathbf{W}$ via: $\mathbf{W} = \beta(\beta\mathbf{S}\mathbf{H}\mathbf{S}^T + \gamma\mathbf{I})^{-1}\mathbf{S}\mathbf{H}\mathbf{F}$;
  *Step*7 : Update $\mathbf{b}$ via: $\mathbf{b} = \frac{(\mathbf{F}^T - \mathbf{W}^T\mathbf{S})\mathbf{1}}{\mathbf{1}^T\mathbf{1}}$;
  *Step*8 : Update $\mathbf{G}$ employing Algorithm 1;
end

$\mathbf{G}$ is $\mathcal{O}(mnc)$. Therefore, we analyze the computation complexity of updating $\mathbf{A}$ and $\mathbf{S}$. The overall computation cost of updating $\mathbf{A}$ and $\mathbf{S}$ is $\mathcal{O}(mn^2)$. Thus, the overall computation complexity of LR-NMF algorithm is $\mathcal{O}(t(mn^2 + mnc))$, where t is the number of iterations. The proof of the convergence using the updating rules of $\mathbf{A}$ and $\mathbf{S}$ is presented in Appendix A.

D. LR-NMF FOR SEMI-SUPERVISED IMAGE CLUSTERING AND CLASSIFICATION

In the previous work of NMFs for image representation, the clustering result is obtained by employing K-means algorithm acting on the nonnegative representation $\mathbf{S}$. During the procedure of matrix factorization, the rank r is setting as the number of clusters c. The procedure of semi-supervised image clustering based on LR-NMF is specified in Algorithm 3. Given the input original image set, few of which are labeled to incorporate the discriminant information into the matrix factorization. Following the semi-supervised learning of LR-NMF, the discriminant information is propagated from labeled images to unlabeled images. Then, the learned nonnegative representations of unlabeled data are used for ultimate clustering.

As for initialization of the basis and coefficient matrices in the proposed algorithm, we adopt the heuristic method to speed up the optimization. Before entering the proposed algorithm, we first employ the K-means technique to cluster the images into r centroids. Then, the basis matrix can be initialized using the r centroids. Each column of the coefficient matrix is a sparse vector whose nonzero element is achieved using the Euclidean distance between the image and the corresponding centroid.

In this work, one of the optimal solutions, $\mathbf{F}_u$, is the label indicator matrix of the unlabeled images at the training phase. To predict the labels of these unlabeled images, it is convenient to adopt $\mathbf{F}_u$ as well as $\mathbf{S}_u^T\mathbf{W} + \mathbf{1}\mathbf{b}^T$ which approaches

Algorithm 3 Procedure of Semi-Supervised Image Clustering Employing LR-NMF

Input: dataset matrix $\mathbf{X} = \{\mathbf{X}_l, \mathbf{X}_u\}$; Label indicator matrix $\mathbf{Y}_l$ for labeled data $\mathbf{X}_l$; Parameters α, β, λ, and γ. Set the reduced rank $r = c$.
Output: Cluster indexes of data points C.
1 Conduct LR-NMF presented in Algorithm 2;
2 Capture the nonnegative representation of unlabeled data $\mathbf{S}_u$ as the feature matrix;
3 Compute the cluster indexes C by gathering the feature points $(\mathbf{S}_u)_j$, where $j = 1, 2, \cdots, u$, into c clusters employing K-means algorithm;

to $\mathbf{F}_u$ in the optimization. However, the unseen data $\mathbf{X}_t \in \mathbb{R}^{d\times 1}$ which is not included in the training set cannot be projected directly using $\mathbf{W}$ and $\mathbf{b}$. Thanks to $\mathbf{A}$ derived from the optimization of LR-NMF, it can extract the nonnegative representation of testing data by $\mathbf{S}_t = \mathbf{A}^T\mathbf{X}_t \in \mathbb{R}^{r\times 1}$ which fits for $\mathbf{W}$ and $\mathbf{b}$. Then, the label can be captured by searching for the index of maximum value of $\mathbf{W}^T\mathbf{S}_t + \mathbf{b}$. The procedure of semi-supervised image classification employing LR-NMF is outlined in Algorithm 4.

IV. EXPERIMENT ANALYSIS

A. EXPERIMENT SETUP

In this section, we conduct the experiment using LR-NMF compared with the competitive NMF algorithms and dimension reduction based semi-supervised learning algorithms. There are eight datasets employed for verifying the competitive algorithms in total:

1) MNIST DIGIT DATASET

The MNIST dataset of handwritten digits has a training set of 60000 examples, and a test set of 10000 examples. It is a subset of a larger set available from NIST. The digits

Algorithm 4 Procedure of Semi-Supervised Image Classification Employing LR-NMF

Input: dataset matrix $\mathbf{X} = \{\mathbf{X}_l, \mathbf{X}_u\}$; Label indicator matrix $\mathbf{Y}_l$ for labeled data $\mathbf{X}_l$; Parameters α, β, λ, and γ; Testing data point $\mathbf{X}_t$.

Output: Labels of unlabeled data points $\mathbf{F}_u$; Label of testing data point $l_{\mathbf{X}_t}$.

1 Conduct LR-NMF presented in Algorithm 2;
2 Predict the labels of unlabeled data using $\mathbf{F}_u$ derived from Algorithm 2, or $\mathbf{F}_u = \mathbf{S}_u^T\mathbf{W} + \mathbf{1}\mathbf{b}^T$;
3 Compute the nonnegative representation of testing data point: $\mathbf{S}_t = \mathbf{A}^T\mathbf{X}_t$;
4 Compute the label vector of testing data point: $\mathbf{F}_t = \mathbf{W}^T\mathbf{S}_t + \mathbf{b}$;
5 Predict the unseen data point via $l_{\mathbf{X}_t} = \arg_j \max(\mathbf{F}_t)$.

have been size-normalized and centered in a 28×28 image. In this experiment, the testing 10000 digit images are used for evaluating the performance of the proposed algorithm.

2) USPS DIGIT DATASET

The USPS dataset consists of a training set with 7291 images and a test set with 2007 images. Each digit image is resized into 16×16 pixels. In order to verify the performance, the test set across ten classes is used in the experiment.

3) ORL FACE DATASET

ORL face image dataset consists of ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions and facial details. All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). The face images used in this work are resized into 32×32 which is similar to [16].

4) YALE FACE DATASET

Yale dataset contains 165 gray scale images of 15 individuals. There are 11 images per subject, one per different facial expression or configuration. For fairly comparing with other algorithms, we also employ the pre-processed Yale dataset in which each image is normalized into 32×32.

5) UMIST FACE DATASET

UMIST face dataset is the last face dataset used in this experiment. The UMIST face dataset consists of 575 images of 20 individuals. Each individual is shown in a range of poses from profile to frontal views. In this experiment, each face image is resized into 28×23 pixels.

6) COIL 20 OBJECT DATASET

COIL 20 dataset is a well-known object dataset and usually used in testifying clustering and classification algorithm. There are totally 1440 images and 72 individuals sharing 20 classes. Each image is normalized into 32×32 resolution.

7) COIL 100 OBJECT DATASET

COIL 100 dataset contains 100 objects using 7200 images in total. The images of each objects were taken 5 degrees apart as the object is rotated on a turntable and each object has 72 images. The size of each image is 32×32 pixels, with 256 grey levels per pixel. Thus, each image is represented by a 1024-dimensional vector.

8) DESCRIBABLE TEXTURES DATASET

Describable textures dataset (DTD) is a texture database [46], consisting of 5640 images, organized according to a list of 47 terms (categories) inspired from human perception. There are 120 images for each category. Image sizes range between 300×300 and 640×640, and the images contain at least 90% of the surface representing the category attribute. The images were collected from Google and Flickr by entering our proposed attributes and related terms as search queries. In this experiment, the deep convolutional activation feature (DeCAF) [47] corresponding to each image is collected for constructing a 4096-dimensional vector. We first reduce the dimension of the DeCAF into 512-dimensional vectors using PCA technique and then normalize them into the nonnegative ones to meet the requirement.

The example images of eight datasets are shown in Fig. 1. For each dataset, we conduct the semi-supervised clustering and classification task according to Algorithm 3 and Algorithm 4. The statics of dataset and the corresponding settings are reported in Table 2. Considering the characters of dataset, we have different settings on number of labeled images and training images for each class. Generally speaking, the number of labeled images is much less than that of unlabeled ones in the real application. For instance, there are about 5 labeled images for each category on the COIL20 dataset.

B. COMPETITIVE ALGORITHM

For evaluating the performance of proposed LR-NMF for semi-supervised image clustering and classification, we employ the semi-supervised NMFs and dimension reduction based semi-supervised learning methods for comparison. The competitive algorithms and the specified settings are as follows:

1) UNSUPERVISED NMFs

As conducted in [16] and [25], we conduct the same procedure which consists of representation learning and K-means for clustering. Thus, K-means is adopted as baseline and NMF [1], [43], ONMF [15], and GNMF [16] are employed as unsupervised NMFs for image clustering. The rank is setting to the number of expected clusters. Since the initialization of K-means influences the performance, we adopt the function *litekmeans* [48] and report the average result of 20 times with different initial points. In addition, the number of nearest neighbors is selected varying from 2 to 5 and then report the best results of GNMF in which $\lambda = 100$. All the algorithms are independently repeated ten times and the average

FIGURE 1. Example images from the MNIST, USPS, ORL, YALE, UMIST, COIL20, COIL100, and DTD datasets from top to bottom. Each row presents 20 images.

TABLE 2. Statics and settings of datasets.

Datasets	Dimension	No. of images	No. of classes	Experimental Settings
MNIST	784	10000	10	60% training and 20% labeled
USPS	256	2007	10	60% training and 20% labeled
ORL	1024	400	40	60% training and 50% labeled
YALE	1024	165	15	60% training and 30% labeled
UMIST	644	575	20	60% training and 30% labeled
COIL20	1024	1440	20	60% training and 12% labeled
COIL100	1024	7200	20	60% training and 12% labeled
DTD	512	5640	47	60% training and 20% labeled

clustering results in terms of the clustering accuracy and normalized mutual information are reported in the later subsection.

2) SEMI-SUPERVISED NMFs

As for the semi-supervised algorithms of NMFs, we employ the semi-supervised NMF (SSNMF) [24], CNMF [25], semi-supervised GNMF (SGNMF) [16], LpNMF [26], and the recently proposed CPSNMF [27] in this experiment for comparison. It is announced that CNMF based on KL divergence performs better than that based on Forbenious norm. Hence, we select KL divergence based CNMF as the competitive method. The λ in SSNMF is set to 1. In the implementation of CPSNMF, $\alpha = 0.2$, $\lambda = 100$.

3) SEMI-SUPERVISED PROJECTION LEARNING

As aforementioned, semi-supervised learning should handle the out-of-sample problem. We adopt Gaussian Field and Harmonic function (GFHF) [32] as baseline algorithm for semi-supervised classification. In addition, SDA [49], SODA [42], and VLR [42] are employed for evaluating the performance on semi-supervised classification. In the implementation of SODA, $\lambda = 20\sigma_{\max}$, where $\sigma_{\max}$ is the maximum diagonal element of S_w defined in the reference. In VLR, α is tuned following to the description in the corresponding paper, $\alpha = 0.9999$ and $\gamma = 1$. All the number of k nearest neighbors is set to the same value with GNMF for fair comparison.

C. SEMI-SUPERVISED IMAGE CLUSTERING

Similar to competitive NMF algorithms, the nonnegative representation derived from LR-NMF denotes the low-dimensional feature points. Given the expected number of clusters which is equal to the dimension of feature, we execute K-means algorithm for ultimate clustering. In this paper, the clustering accuracy (ACC) and normalized mutual information (NMI) are used for evaluating the performance of the competitive algorithms.

ACC is computed by:

$$ACC = \frac{\sum_{i=1}^{n} \phi(l_i, bestmap(k_i))}{n}, \tag{38}$$

where l_i and k_i are the true and predicted label of $\mathbf{X}_i$, respectively, and the indicator function $\phi(a, b)$ equals 1 if $a = b$; 0, otherwise. Furthermore, The function $bestmap(\cdot)$ matches the true label and the predicted label and the best mapping is solved by Kuhn-Munkres algorithm.

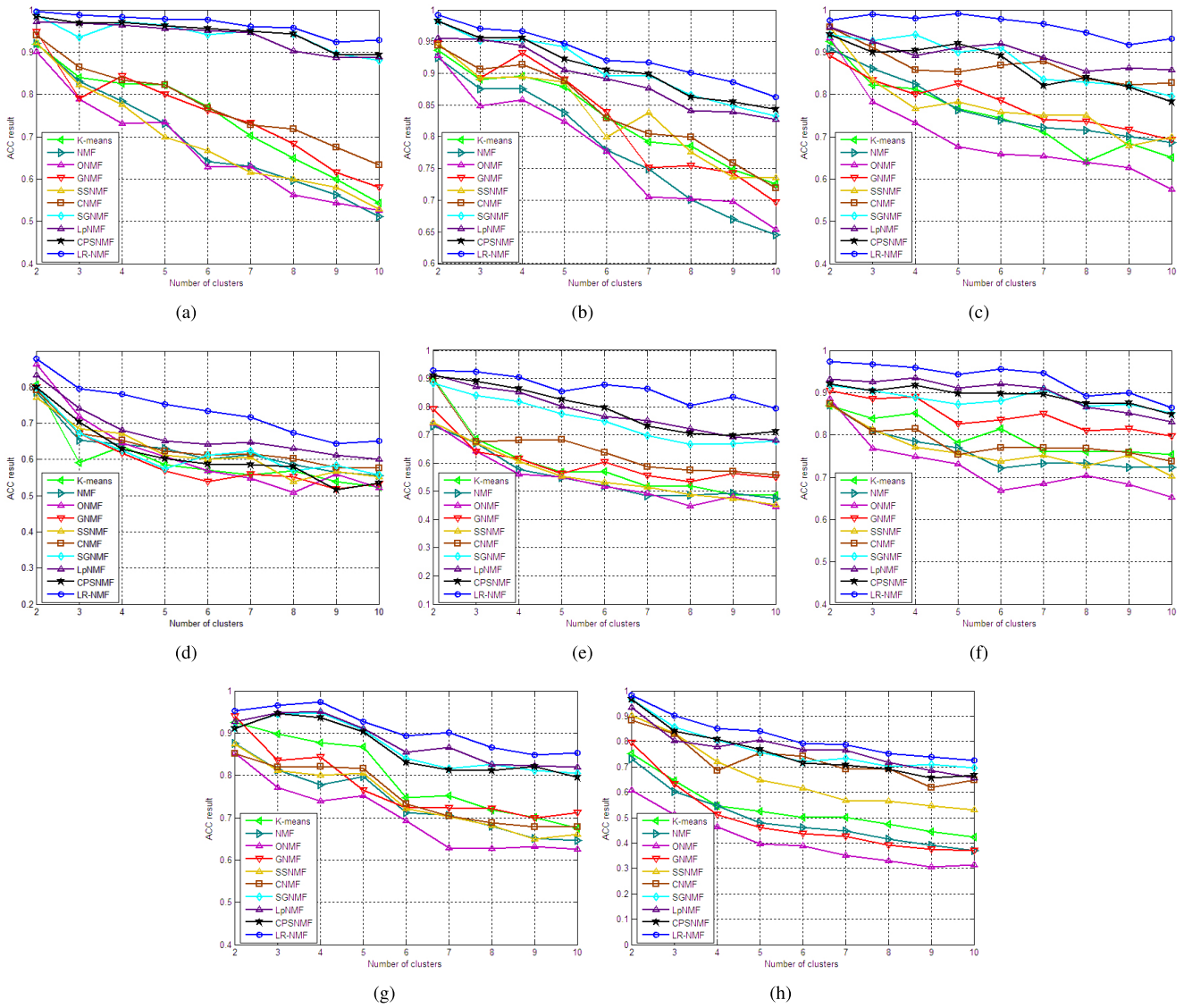


FIGURE 2. ACC curves of competitive algorithms versus different numbers of clusters on the (a) MNIST, (b) USPS, (c) ORL, (d) YALE, (e) UMIST, (f) COIL20, (g) COIL100, and (h) DTD datasets.

NMI is define as follows:

$$NMI(A, B) = \frac{I(A, B)}{\sqrt{H(A)H(B)}}, \tag{39}$$

where $I(A, B)$ is the mutual information between A and B, and $H(A)$ and $H(B)$ are the entropy of A and B, respectively. Both ACC and NMI take a value within the interval [0,1]. The higher the values, the better the clustering performance.

For varying the performance of clustering, we run the algorithms with different given numbers of clusters varying from 2 to 10. With a fixed number of clusters, the images are selected from each dataset in the light of settings reported in Table 2. The curves of ACC and NMI using the competitive algorithms on the employed datasets are shown in Fig. 2 and Fig. 3, respectively.

From Fig. 2 and Fig. 3, it can be observed some interested points:

1) The proposed LR-NMF achieves the best clustering results with different required numbers of clusters on the challenging datasets. Observed by the clustering results in terms of ACC and NMI, LR-NMF improves largely the performance on image representation. It illustrates that LR-NMF imposes more discriminative information employing label propagation during the procedure of NMF.
2) Most of the semi-supervised NMFs outperforms the unsupervised ones except that SSNMF performs worse than others. Among the unsupervised ones, GNMF gets impressive results due to preserving the structure of data.
3) As for the clustering results on the UMIST and COIL20 datasets, the competitive CPSNMF algorithm performs better than that on the other datasets. Typically, it even slightly approaches to the

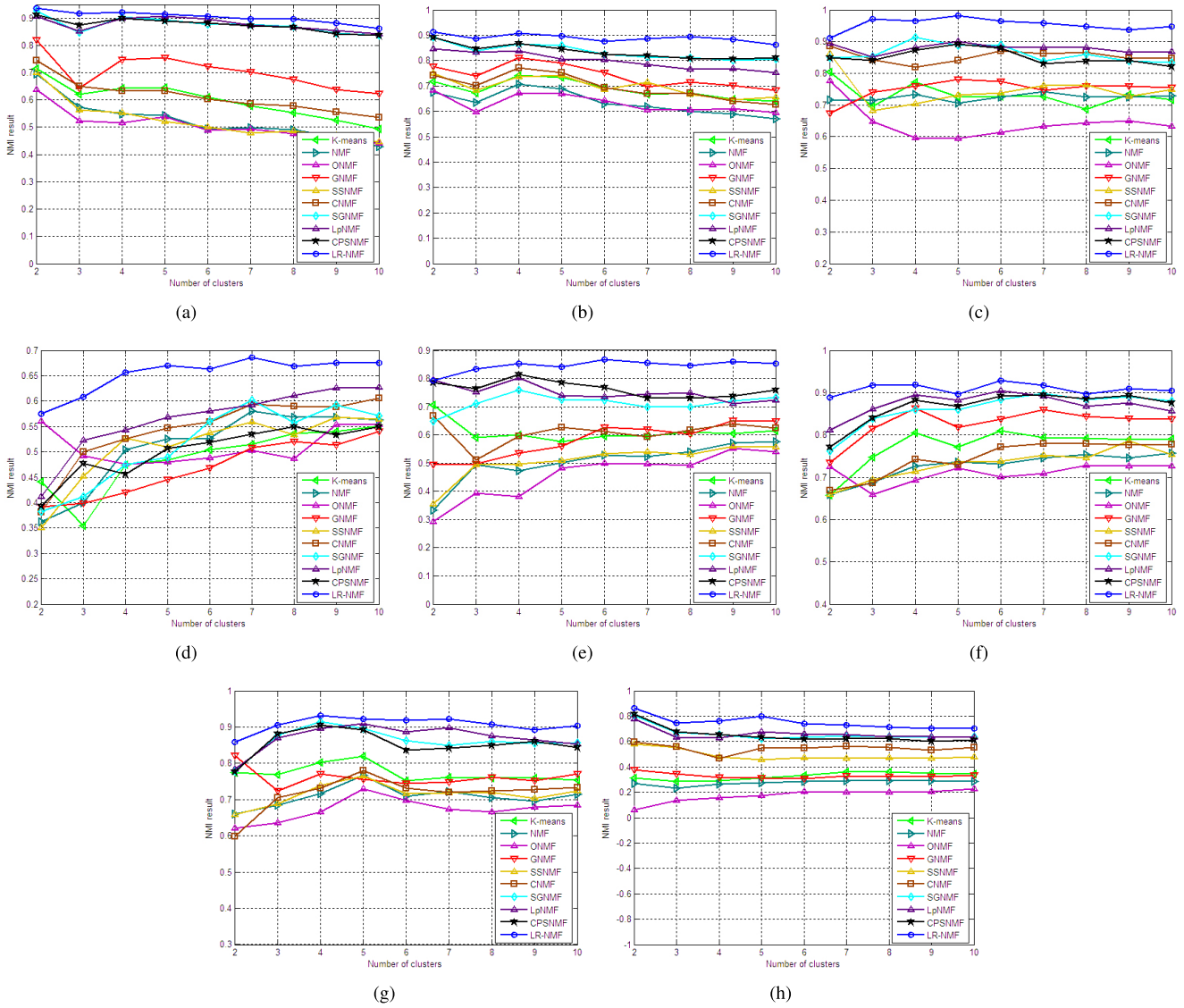


FIGURE 3. NMI curves of competitive algorithms versus different numbers of clusters on the (a) MNIST, (b) USPS, (c) ORL, (d) YALE, (e) UMIST, (f) COIL20, (g) COIL100, and (h) DTD datasets.

proposed LR-NMF. However, it can be obviously observed that LR-NMF outperforms CPSNMF in most cases.

4) The LpNMF algorithm outperforms the other compared NMF variants on the MNIST dataset. On the other datasets, LpNMF achieves similar results with CPSNMF. The performance comparison results of LpNMF, CPSNMF, and SGNMF depend on the number of clusters. This can be observed on the USPS, UMIST, and COIL20 datasets.
5) Among the competitive algorithms, CNMF, SGNMF, LpNMF, and CPSNMF which are the second echelon achieve better clustering results than other compared algorithms. Since CPSNMF and SGNMF are conducted based on the framework of GNMF, they get the similar results. However, CPSNMF outperforms slightly SGNMF in some cases.

D. SEMI-SUPERVISED IMAGE CLASSIFICATION

For predicting the unseen images which are not included in the training set, LR-NMF learns a pair of parameters of linear regression on the nonnegative representation. According to Algorithm 3, we run the proposed algorithm by repeating five times with different splits and report the classification results of the compared algorithms on the eight datasets in Table 3. Since GFHF cannot predict the unseen data, we only report the results of predicting unlabeled data in Table 3. From Table 3, LR-NMF gets the best classification result on the eight datasets. It illustrates that the proposed LR-NMF can also performs well in semi-supervised classification problem.

From the compared results of semi-supervised clustering and classification algorithms, it can be seen that LR-NMF improves the performance for learning the image representation. Moreover, the products of LR-NMF can also be

TABLE 3. Semi-supervised classification results of unlabeled and unseen images using competitive algorithms on the eight datasets (mean ± standard deviation)%.

dataset	Testing Images	GFHF	SDA	SODA	VLR	LR-NMF
MNIST	Unlabeled	88.98 ± 0.50	84.23± 0.65	84.58 ± 0.49	86.55± 0.65	91.05 ± 0.62
	Unseen	unavailable	83.10± 0.51	84.01 ± 0.36	82.09± 0.54	86.22 ± 0.43
USPS	Unlabeled	86.71 ± 1.37	82.62± 1.70	81.79 ± 1.65	84.38 ± 0.90	88.72 ± 1.38
	Unseen	unavailable	83.57± 1.15	82.82 ± 0.74	82.86 ± 2.19	85.85 ± 1.07
ORL	Unlabeled	75.67 ± 3.70	79.67± 5.29	78.83 ± 5.93	80.33 ± 5.42	86.32 ± 4.08
	Unseen	unavailable	85.50± 1.08	82.88 ± 2.33	85.75 ± 0.73	87.96 ± 1.24
YALE	Unlabeled	48.27 ± 5.68	49.87± 5.76	53.33 ± 7.11	49.87 ± 5.76	57.85 ± 7.26
	Unseen	unavailable	53.67± 3.23	49.67 ± 8.06	56.67 ± 3.50	57.68 ± 4.78
UMIST	Unlabeled	85.76 ± 3.29	89.81 ± 2.45	87.78 ± 1.85	88.09 ± 1.17	91.14 ± 1.44
	Unseen	unavailable	89.64± 3.81	87.75 ± 4.01	87.48 ± 4.68	90.15 ± 3.52
COIL20	Unlabeled	88.36 ± 1.98	87.56 ± 1.65	87.03 ± 2.14	88.97 ± 1.39	89.75 ± 1.37
	Unseen	unavailable	84.43 ± 2.21	85.25 ± 3.86	85.39 ± 2.14	86.69 ± 3.22
COIL100	Unlabeled	82.96 ± 2.16	69.61 ± 3.50	77.72 ± 3.41	81.45 ± 2.18	85.43 ± 2.68
	Unseen	unavailable	65.51 ± 6.09	75.25 ± 6.55	71.53 ± 2.30	78.49 ± 3.81
DTD	Unlabeled	46.29 ± 0.48	41.42 ± 1.40	41.37 ± 0.80	48.22 ± 0.66	49.88 ± 0.57
	Unseen	unavailable	40.57 ± 0.86	41.10 ± 0.96	45.58 ± 0.90	47.62 ± 0.72

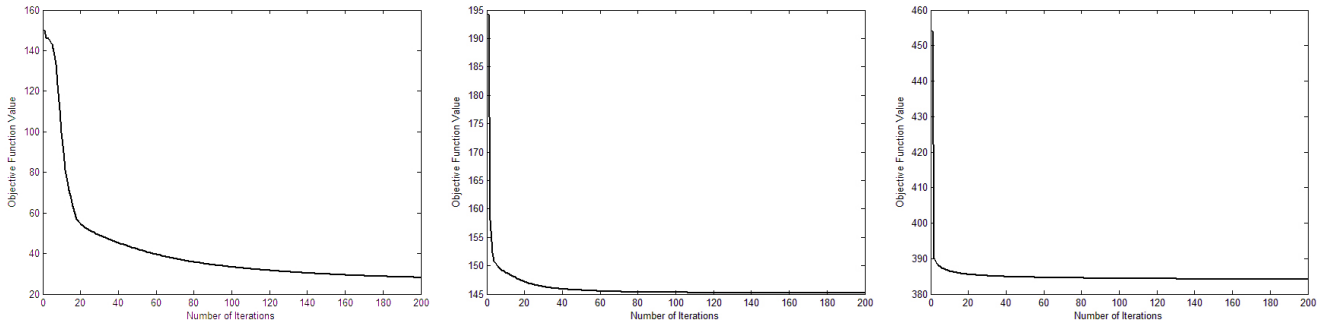


FIGURE 4. Convergence curves of proposed LR-NMF algorithm on the (a) MNIST, (b) UMIST, and (c) COIL20 datasets.

employed for predicting the labels of unlabeled and unseen data.

E. CONVERGENCE STUDY

It has been proved that the objective of LR-NMF obtains local optimums due to updating the basis matrix and the corresponding representation. We have recorded the value of cost function versus the number of iterations in the experiment. In order to illustrate the decreasing of the objective during the updating processing, we select three datasets: MNIST digit, UMIST face, and COIL20 object datasets to study the convergence of the proposed LR-NMF algorithm. The convergence curves associated with the number of iterations are shown in Fig. 4. From Fig. 4, it can be observed that the curves drop rapidly on the datasets at the beginning phase of learning. Thus, the convergence of the proposed algorithm on the datasets guarantees the implementation of the method for semi-supervised learning.

F. PARAMETERS SELECTION

There are two groups of parameters in the proposed algorithm. One of the groups lies in Algorithm 1, including the

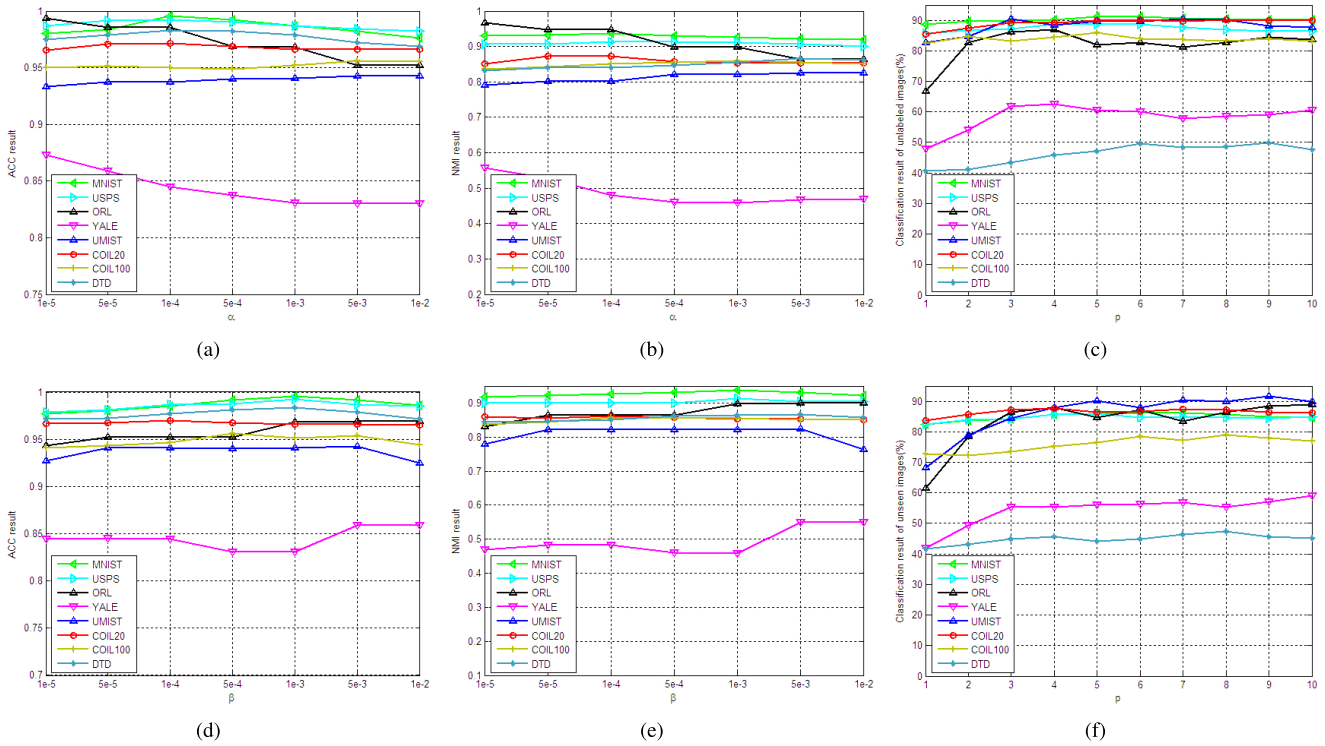


FIGURE 5. The clustering results in terms of ACC v.s. α (a) and β (d) and NMI v.s. α (b) and β (e), and the classification results of unlabeled images (c) and unseen images (f) v.s. p on the eight datasets.

learning step η, λ, and the sparsity K, to accomplish the self-organized graph learning. Actually, this problem can be conducted by setting $\eta = 0.01$, $\lambda = 0.005$, and $K = 3$ on the face datasets and $K = 5$ on the other datasets according to the pattern complexity of the images. It should be noticed that the graph learning is not the ultimate goal, we only set the parameters simply to get a valid graph during the experiment. Another group of parameters consists of α, β, γ, and r weight the importances of the terms in the objective. In order to decrease the cost of selecting the parameters, γ is set as 10^{-4} due to its role of regularization. Then, we select α, β, and r from the the cube of $\{10^{-5}, 5 \times 10^{-5}, 10^{-4}, \cdots, 10^{-2}\}$, $\{10^{-5}, 5 \times 10^{-5}, 10^{-4}, \cdots, 10^{-2}\}$, and $\{c, 2c, 3c, \cdots, 10c\}$, where c is the number of classes. Then, we conduct the LR-NMF algorithm by randomly selecting two classes based on the setting in Table 3 and report the average clustering results in terms of ACC and NMI on the eight datasets. The performance results versus α with $\beta = 10^{-3}$ and β with $\alpha = 10^{-3}$ are shown in the first two columns of Fig. 5. In this experiment, we set $\alpha = 10^{-3}$, $\beta = 10^{-3}$ in a nutshell. For semi-supervised classification task, the dimension of nonnegative representation influences the performance. In order to analyze the effect to classification results, we also show the results of unlabeled and unseen images versus the value of $r = pc$ in the last column of Fig. 5, where p is varying from 1 to 10. In this experiment, we simply set the dimension to $5c$ on the UMIST dataset and $3c$ on other datasets which make the dimension be smaller than the number of training images. From Fig. 5, we can see that the results of clustering vary slightly and those of classification achieve better performance with relatively larger dimension.

V. CONCLUSION

This paper proposes a fully functional NMF approach to nonnegative feature extraction and linear regression in semi-supervised fashion. By incorporating the linear regression and label propagation based on the self-organized graph into the optimization of NMF, the proposed LR-NMF can jointly spread the supervised information to the unlabeled data during the matrix decomposition and conduct the linear regression acting on the nonnegative feature. Furthermore, the alternatively updating rules of the variables are given and proved to converge to local optimums. Extensive experiments on the eight challenging datasets for semi-supervised image clustering verify that the proposed LR-NMF improves significantly the clustering performance compared with the state-of-the-art semi-supervised NMF algorithms for image representation. Meanwhile, LR-NMF achieves better results for classification task of unlabeled and unseen data compared with the competitive semi-supervised learning algorithms. Although the basis matrix has been constrained orthogonally, it remains challenging to exactly get the nonnegative representation for predicting the unseen data. The transfer learning method for solving this problem will be discussed in the future work.

APPENDIX
PROOF OF THE CONVERGENCE OF LR-NMF FOR UPDATING RULES OF A AND S

In order to prove the convergence of updating rules of $\mathbf{A}$ and $\mathbf{S}$, we introduce the auxiliary function defined using Expectation-Maximization algorithm [50], [51] as follows:

Definition 1: Function $g(x, x^t)$ is an auxiliary function for $f(x)$ if the conditions $g(x, x^t) \geq f(x)$, $g(x, x) = f(x)$ are satisfied.

According to Theorem 1, updating rules of $\mathbf{A}$ and $\mathbf{S}$ make the alternative update process converge to local optimum. Lemma 1 proves $f(x)$ is nonincreasing.

Theorem 1: The objective functions in (13) and (14) are nonincreasing under the updating rules in (23) and (24). The objective function w.r.t. $\mathbf{A}$ and $\mathbf{S}$ are invariant under these updates if and only if $\mathbf{A}$ and $\mathbf{S}$ are at the stationary points, respectively.

Lemma 2: If g is an auxiliary function of f, then f is nonincreasing under the update $x^{t+1} = \arg\min_x g(x, x^t)$.

Without loss of generality, $\mathbf{A}_{ij}$ and $\mathbf{S}_{ij}$ denote any element of $\mathbf{A}$ and $\mathbf{S}$, $J_{\mathbf{A}}$ and $J_{\mathbf{S}}$ is the part of the objective relevant to $\mathbf{S}$. Let $\Delta\mathbf{X} = \mathbf{X} - \mathbf{X}\mathbf{G}$, we define the auxiliaries of $J_{\mathbf{A}}$ and $J_{\mathbf{S}}$ as follows:

$$g(\mathbf{A}_{ij}, \mathbf{A}_{ij}^t) = J_{\mathbf{A}}(\mathbf{A}_{ij}^t) + 2(\mathbf{A}\mathbf{S}\mathbf{S}^T + \lambda\Delta\mathbf{X}\Delta\mathbf{X}^T\mathbf{A})_{ij}(\mathbf{A}_{ij} - \mathbf{A}_{ij}^t), \tag{40}$$

and

$$g(\mathbf{S}_{ij}, \mathbf{S}_{ij}^t) = J_{\mathbf{S}}(\mathbf{S}_{ij}^t) + J_{\mathbf{S}}^{(1)}(\mathbf{S}_{ij}^t)(\mathbf{S}_{ij} - \mathbf{S}_{ij}^t) + \frac{(\mathbf{A}^T\mathbf{A}\mathbf{S} + \beta\mathbf{W}\mathbf{W}^T\mathbf{S})_{ij}}{\mathbf{S}_{ij}^t}(\mathbf{S}_{ij} - \mathbf{S}_{ij}^t)^2. \tag{41}$$

Proof: It is obvious that $g(\mathbf{A}_{ij}, \mathbf{A}_{ij}) = J_{\mathbf{A}}(\mathbf{A}_{ij})$ and $g(\mathbf{S}_{ij}, \mathbf{S}_{ij}) = J_{\mathbf{S}}(\mathbf{S}_{ij})$. In order to prove $g(\mathbf{A}_{ij}, \mathbf{A}_{ij}^t) \geqslant J_{\mathbf{A}}(\mathbf{A}_{ij})$ and $g(\mathbf{S}_{ij}, \mathbf{S}_{ij}^t) \geqslant J_{\mathbf{S}}(\mathbf{S}_{ij})$, we first express the Taylor series expansions of $J_{\mathbf{A}}(\mathbf{A}_{ij})$ and $J_{\mathbf{S}}(\mathbf{S}_{ij})$ as:

$$J_{\mathbf{A}}(\mathbf{A}_{ij}) = J_{\mathbf{A}}(\mathbf{A}_{ij}^t) + J_{\mathbf{A}}^{(1)}(\mathbf{A}_{ij}^t)(\mathbf{A}_{ij} - \mathbf{A}_{ij}^t), \tag{42}$$

and

$$J_{\mathbf{S}}(\mathbf{S}_{ij}) = J_{\mathbf{S}}(\mathbf{S}_{ij}^t) + J_{\mathbf{S}}^{(1)}(\mathbf{S}_{ij}^t)(\mathbf{S}_{ij} - \mathbf{S}_{ij}^t) + \frac{J_{\mathbf{S}}^{(2)}(\mathbf{S}_{ij}^t)}{2}(\mathbf{S}_{ij} - \mathbf{S}_{ij}^t)^2. \tag{43}$$

Then, we compare (42) and (43) with the auxiliary functions in (40) and (41).

It's well checked that:

$$\begin{aligned} J_{\mathbf{A}}^{(1)} &= 2\mathbf{A}\mathbf{S}\mathbf{S}^T + 2\lambda\Delta\mathbf{X}\Delta\mathbf{X}^T\mathbf{A}, \\ J_{\mathbf{S}}^{(1)} &= 2\mathbf{A}^T(\mathbf{A}\mathbf{S} - \mathbf{X}) + 2\beta\mathbf{W}\mathbf{W}^T S + 2\beta\mathbf{W}(\mathbf{b}\mathbf{1}^T - \mathbf{F}^T), \\ J_{\mathbf{S}}^{(2)} &= 2\mathbf{A}^T\mathbf{A} + 2\beta\mathbf{W}\mathbf{W}^T. \end{aligned} \tag{44}$$

To prove $g(\mathbf{A}_{ij}, \mathbf{A}_{ij}^t) \geqslant J_{\mathbf{A}}(\mathbf{A}_{ij})$ and $g(\mathbf{S}_{ij}, \mathbf{S}_{ij}^t) \geqslant J_{\mathbf{S}}(\mathbf{S}_{ij})$, it is equivalent to prove the following inequalities:

$$(\mathbf{A}\mathbf{S}\mathbf{S}^T + \lambda\Delta\mathbf{X}\Delta\mathbf{X}^T\mathbf{A})_{ij} \geqslant \mathbf{A}_{ij}^t\mathbf{S}\mathbf{S}_{ij}^T + \lambda(\Delta\mathbf{X}\Delta\mathbf{X}^T)_{ij}\mathbf{A}_{ij}^t, \tag{45}$$

and

$$\frac{(\mathbf{A}^T\mathbf{A}\mathbf{S} + \beta\mathbf{W}\mathbf{W}^T\mathbf{S})_{ij}}{\mathbf{S}_{ij}^t} \geqslant \frac{J_{\mathbf{S}}^{(2)}(\mathbf{S}_{ij}^t)}{2},$$
$$(\mathbf{A}^T\mathbf{A}\mathbf{S} + \beta\mathbf{W}\mathbf{W}^T\mathbf{S})_{ij} \geqslant (\mathbf{A}^T\mathbf{A} + \beta\mathbf{W}\mathbf{W}^T)_{ij}\mathbf{S}_{ij}^t. \tag{46}$$

Thus $g(\mathbf{A}_{ij}, \mathbf{A}_{ij}^t) \geqslant J_{\mathbf{A}}(\mathbf{A}_{ij})$ and $g(\mathbf{S}_{ij}, \mathbf{S}_{ij}^t) \geqslant J_{\mathbf{S}}(\mathbf{S}_{ij})$ hold.

Denote the value of objective function at the t-th iteration by $J(\mathbf{W}^t, \mathbf{b}^t, \mathbf{F}_u^t, \mathbf{G}^t, \mathbf{A}^t, \mathbf{S}^t)$. While updating one of the variables $\{\mathbf{W}, \mathbf{b}, \mathbf{F}_u, \mathbf{G}\}$ and fixing others, the corresponding function is smooth and derivable, leading to convergence to global optimum. Based on the closed solutions of $\{\mathbf{W}, \mathbf{b}, \mathbf{F}_u, \mathbf{G}\}$ and the convergence proof of updating $\mathbf{A}$ and $\mathbf{S}$, we have:

$$\begin{aligned} &J(\mathbf{W}^{t+1}, \mathbf{b}^{t+1}, \mathbf{F}_u^{t+1}, \mathbf{G}^{t+1}, \mathbf{A}^{t+1}, \mathbf{S}^{t+1}) \\ &\leq J(\mathbf{W}^t, \mathbf{b}^t, \mathbf{F}_u^t, \mathbf{G}^t, \mathbf{A}^{t+1}, \mathbf{S}^{t+1}) \\ &\leq J(\mathbf{W}^t, \mathbf{b}^t, \mathbf{F}_u^t, \mathbf{G}^t, \mathbf{A}^t, \mathbf{S}^t) \end{aligned} \tag{47}$$

Therefore, it has been proved that the updating rules of the variables involved in the objective of proposed LR-NMF guarantee the convergence to local optimums.

REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] I. Buciu, N. Nikolaidis, and I. Pitas, "Nonnegative matrix factorization in polynomial feature space," *IEEE Trans. Neural Netw.*, vol. 19, no. 6, pp. 1090–1100, Jun. 2008.

[3] Y. Chen, J. Zhang, D. Cai, W. Liu, and X. He, "Nonnegative local coordinate factorization for image representation," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 969–979, Mar. 2013.

[4] Z. Wang, X. Kong, H. Fu, M. Li, and Y. Zhang, "Feature extraction via multi-view non-negative matrix factorization with local graph regularization," in *Proc. Int. Conf. Image Process.*, Sep. 2015, pp. 3500–3504.

[5] Z. Li, J. Tang, and X. He, "Robust structured nonnegative matrix factorization for image representation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1947–1960, May 2018.

[6] C. Peng, Z. Kang, Y. Hu, J. Cheng, and Q. Cheng, "Nonnegative matrix factorization with integrated graph and feature learning," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, p. 42, 2017.

[7] Q. Qu, N. M. Nasrabadi, and T. D. Tran, "Subspace vertex pursuit: A fast and robust near-separable nonnegative matrix factorization method for hyperspectral unmixing," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1142–1155, Sep. 2015.

[8] X. Wang, Y. Zhong, L. Zhang, and Y. Xu, "Spatial group sparsity regularized nonnegative matrix factorization for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6287–6304, Nov. 2017.

[9] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 38–52, Feb. 2011.

[10] S. Nikitidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Subclass discriminant nonnegative matrix factorization for facial image analysis," *Pattern Recognit.*, vol. 45, no. 12, pp. 4080–4091, 2012.

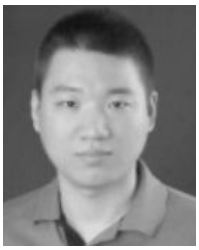
[11] S. Essid and C. Févotte, "Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 415–425, Feb. 2013.

[12] X. Pei, T. Wu, and C. Chen, "Automated graph regularized projective nonnegative matrix factorization for document clustering," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1821–1831, Oct. 2014.

[13] Z. Yuan and E. Oja, "Projective nonnegative matrix factorization for image compression and feature extraction," in *Proc. Scandin. Conf. Image Anal.*, 2005, pp. 333–342.

[14] C. Ding, T. Li, W. Peng, and T. Li, "Orthogonal nonnegative matrix tri-factorizations for clustering," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 126–135.

[15] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 1828–1832.
[16] D. Cai, X. He, J. Han, and T. Huang, "Graph regularized non-negative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.
[17] C. Peng, Z. Kang, Y. Hu, and Q. Cheng, "Robust graph regularized nonnegative matrix factorization for clustering," *ACM Trans. Knowl. Discovery Data*, vol. 11, no. 3, pp. 1–21, 2017.
[18] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE Trans. Neural Netw.*, vol. 17, no. 3, pp. 683–695, May 2006.
[19] I. Kotsia, S. Zafeiriou, and I. Pitas, "A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 588–595, Sep. 2007.
[20] S. Zafeiriou, "Discriminant nonnegative tensor factorization algorithms," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 217–235, Feb. 2009.
[21] A. Vilamala, P. J. G. Lisboa, S. Ortega-Martorell, and A. Vellido, "Discriminant convex non-negative matrix factorization for the classification of human brain tumours," *Pattern Recognit. Lett.*, vol. 34, no. 14, pp. 1734–1747, 2013.
[22] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 2030–2048, Jul. 2011.
[23] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, and C. Yuan, "Nonnegative discriminant matrix factorization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 7, pp. 1392–1405, Jul. 2017.
[24] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *IEEE Signal Process. Lett.*, vol. 17, no. 1, pp. 4–7, Jan. 2010.
[25] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Huang, "Constrained nonnegative matrix factorization for image representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1299–1311, Jul. 2012.
[26] Y. Yi, Y. Shi, H. Zhang, J. Wang, and J. Kong, "Label propagation based semi-supervised non-negative matrix factorization for feature extraction," *Neurocomputing*, vol. 149, pp. 1021–1037, Feb. 2015.
[27] D. Wang, X. Gao, and X. Wang, "Semi-supervised nonnegative matrix factorization via constraint propagation," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 233–244, Jan. 2016.
[28] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[29] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based Fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 209–232, Sep. 2014.
[30] W. Zhu, Y. Yan, and Y. Peng, "Dictionary learning based on discriminative energy contribution for image classification," *Knowl. Based Syst.*, vol. 113, pp. 116–124, Dec. 2016.
[31] C. Zhang, H. Zheng, and J. Lai, "Cross-view action recognition based on hierarchical view-shared dictionary learning," *IEEE Access*, vol. 6, pp. 16855–16868, 2018.
[32] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. Int. Conf. Mach. Learn.*, 2003, pp. 912–919.
[33] F. Nie, S. Xiang, Y. Jia, and C. Zhang, "Semi-supervised orthogonal discriminant analysis via label propagation," *Pattern Recognit.*, vol. 42, no. 11, pp. 2615–2627, 2009.
[34] X. Wang, Y. Liu, F. Nie, and H. Huang, "Discriminative unsupervised dimensionality reduction," in *Proc. Int. Conf. Artif. Intell.*, 2015, pp. 3925–3931.
[35] R. Wang, F. Nie, R. Hong, X. Chang, X. Yang, and W. Yu, "Fast and orthogonal locality preserving projections for dimensionality reduction," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 5019–5030, Oct. 2017.
[36] W. Zhu, Y. Yan, and Y. Peng, "Pair of projections based on sparse consistence with applications to efficient face recognition," *Signal Process. Image Commun.*, vol. 55, pp. 32–40, Jul. 2017.
[37] M. Mei, J. Huang, and W. Xiong, "A discriminant subspace learning based face recognition method," *IEEE Access*, vol. 6, pp. 13050–13056, 2017.
[38] C. Peng, J. Cheng, and Q. Cheng, "A supervised learning model for high-dimensional and large-scale data," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 2, p. 30, 2017.
[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[40] S. Xie, T. Yang, X. Wang, and Y. Lin, "Hyper-class augmented and regularized deep learning for fine-grained image classification," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2645–2654.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[42] F. Nie, D. Xu, X. Li, and S. Xiang, "Semisupervised dimensionality reduction and classification through virtual label regression," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 3, pp. 675–685, Jun. 2011.
[43] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2000, pp. 535–541.
[44] F. Nie, H. Wang, H. Huang, and C. Ding, "Adaptive loss minimization for semi-supervised elastic embedding," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1565–1571.
[45] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.
[46] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3606–3613.
[47] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. IEEE Conf. Mach. Learn.*, Jan. 2014, pp. 647–655.
[48] D. Cai, X. He, X. Wu, and J. Han, "Non-negative matrix factorization on manifold," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 63–72.
[49] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–7.
[50] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B, Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.
[51] L. Saul and F. Pereira, "Aggregate and mixed-order Markov models for statistical language processing," in *Proc. 2nd Conf. Empirical Methods Natural Lang. Process.*, Jun. 1997, pp. 81–89.



WENJIE ZHU received the B.S. degree from the School of Science, North China University of Science and Technology, in 2010, and the M.S. degree from the School of Electronic Engineering, Xidian University, in 2013. He is currently pursuing the Ph.D. degree with the School of Mechanical Engineering and Automation, Northeastern University, Shenyang, China. His research interests include machine learning and pattern recognition.

YUNHUI YAN received the B.S., M.S., and Ph.D. degrees from the School of Mechanical Engineering and Automation, Northeastern University, Shenyang, China, in 1981, 1985, and 1997, respectively. He has been a Teacher with Northeastern University since 1982, and became a Professor in 1997. From 1993 to 1994, he was with the Tohoku National Industrial Research Institute as a Visiting Scholar. His research interest covers intelligent inspection, image processing, and pattern recognition.