Hindawi
BioMed Research International
Volume 2017, Article ID 1096028, 8 pages
https://doi.org/10.1155/2017/1096028

Research Article

Gene Feature Extraction Based on Nonnegative Dual Graph Regularized Latent Low-Rank Representation

Guoliang Yang and Zhengwei Hu

School of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou 341000, China

Correspondence should be addressed to Guoliang Yang; ygliang30@126.com and Zhengwei Hu; huzhengwei1993@163.com

Received 21 January 2017; Accepted 13 March 2017; Published 30 March 2017

Academic Editor: Gang Liu

Aiming at the problem of gene expression profile's high redundancy and heavy noise, a new feature extraction model based on nonnegative dual graph regularized latent low-rank representation (NNDGLLRR) is presented on the basis of latent low-rank representation (Lat-LRR). By introducing dual graph manifold regularized constraint, the NNDGLLRR can keep the internal spatial structure of the original data effectively and improve the final clustering accuracy while segmenting the subspace. The introduction of nonnegative constraints makes the computation with some sparsity, which enhances the robustness of the algorithm. Different from Lat-LRR, a new solution model is adopted to simplify the computational complexity. The experimental results show that the proposed algorithm has good feature extraction performance for the heavy redundancy and noise gene expression profile, which, compared with LRR and Lat-LRR, can achieve better clustering accuracy.

1. Introduction

With the accelerated pace of modern life, the high incidence of cancer has brought great challenges to human health. How to detect, prevent, and treat cancer effectively has become an international hotspot of medical research. Gene expression profile is a specific cDNA sequence data of cells, which can describe cells' current physiological function and state. Researches show that tumor cells and normal cells could be identified effectively by analyzing and processing the original gene expression data. However, the scale of the gene expression profile is huge and complex due to the diversity and specificity of the cells; therefore the traditional methods of data analysis and processing have been unable to adapt to these extremely large-scale data.

Gene expression profile extracting includes two kinds of methods: linear and nonlinear. Early linear transformation methods include principal component analysis [1–3] (PCA), linear discriminant analysis [4–6] (LDA), and independent component analysis [7, 8] (ICA). The main methods of nonlinear transformation include kernel method [9], neural network [10, 11], manifold learning [12, 13], and sparse representation [14, 15]. In recent years, LRR [16–18] and neural networks have been widely used in feature extraction and classification of gene expression profile. Reference [19] used NMF for gene feature extraction and achieved more satisfactory results. Ref. [20] proposed a gene expression profile classification means based on ontology perception. Ref. [21] proposed a subcellular cooccurrence matrix feature extraction method. Ref. [22] proposed a gene expression profile classification method by neural network hybrid back-propagation. Ref. [23] proposed a supervised way of tumor prediction with multiview.

The size of the gene expression profile is large, and there are interrelationships between the samples. The internal spatial structure of the data may be destroyed in the process of linear transformation. In this paper, a model of feature extraction based on NNDGLLRR is proposed on the basis of Lat-LRR, which with low-rank sparse constraint can remove the redundant components of gene expression and suppress the noise. Nonnegative constraints make the calculation with a certain degree of sparsity, in line with the practical significance of the data, and enhance the robustness of the algorithm. And the manifold regularized constraint is introduced, so that the result of feature extraction can describe the spatial structure of the original data more completely.

2. Related Work

2.1. LRR. LRR is a combination of matrix low-rank decomposition and sparse decomposition. In recent years, it has been widely used in subspace clustering. LRR assumes that the original data comes from different subspaces and performs feature extraction by trying to find the lowest rank representation of the original data. And this low-rank representation coefficient is the reflection of the original data in the spatial distribution of structural information. If the original data $\mathbf{X} = [x_1, x_2, x_3, \ldots, x_n] \in \mathbb{R}^{m\times n}$, each column x_i represents a sample, and generally the LRR uses the data itself as a dictionary. Then the model can be as shown in

$$O_1 = \begin{cases} \min & \|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_{2,1} \\ \text{s.t.} & \mathbf{X} = \mathbf{XZ} + \mathbf{E}. \end{cases} \tag{1}$$

The LRR matrix $\mathbf{Z} = [z_1, z_2, z_3, \ldots, z_n] \in \mathbb{R}^{n\times n}$, and z_i is the linear representation coefficient of the sample x_i under the data dictionary $\mathbf{X}$. The original data usually contains a lot of noise, while the sparse constraint can maintain the robustness of the algorithm effectively. Ref. [24] shows the specific solution process of LRR.

Let $\mathbf{Z} = \mathbf{J}$; we construct the following Augmented Lagrangian function:

$$\begin{aligned} \mathscr{L} = {} & \|\mathbf{J}\|_* + \lambda \|\mathbf{E}\|_{2,1} + \langle \mathbf{\Lambda}, \mathbf{Z} - \mathbf{J} \rangle + \langle \mathbf{\Pi}, \mathbf{X} - \mathbf{XZ} - \mathbf{E} \rangle \\ & + \frac{\mu}{2} \left\{ \|\mathbf{Z} - \mathbf{J}\|_F^2 + \|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_F^2 \right\}. \end{aligned} \tag{2}$$

The specific update algorithm is as follows.

Keep $\mathbf{Z} = \mathbf{Z}^k, \mathbf{\Lambda} = \mathbf{\Lambda}^k$; update $\mathbf{J}$:

$$\mathbf{J}^{k+1} = \arg\min_{\mathbf{J}} \left\| \mathbf{J}^k \right\|_* + \frac{\mu^k}{2} \left\| \frac{\mathbf{\Lambda}^k}{\mu^k} + \mathbf{Z}^k - \mathbf{J}^k \right\|_F^2. \tag{3}$$

Keep $\mathbf{J} = \mathbf{J}^k$, $\mathbf{\Lambda} = \mathbf{\Lambda}^k$, and $\mathbf{\Pi} = \mathbf{\Pi}^k$; update $\mathbf{Z}$:

$$\begin{aligned} \mathbf{Z}^{k+1} = \arg\min_{\mathbf{Z}} & \left\| \frac{\mathbf{\Lambda}^k}{\mu^k} + \mathbf{Z}^k - \mathbf{J}^k \right\|_F^2 \\ & + \left\| \frac{\mathbf{\Pi}^k}{\mu^k} + \mathbf{X} - \mathbf{XZ}^k - \mathbf{E}^k \right\|_F^2. \end{aligned} \tag{4}$$

Keep $\mathbf{Z} = \mathbf{Z}^k, \mathbf{\Pi} = \mathbf{\Pi}^k$; update $\mathbf{E}$:

$$\begin{aligned} \mathbf{E}^{k+1} = \arg\min_{\mathbf{E}} & \, \lambda \left\| \mathbf{E}^k \right\|_{2,1} \\ & + \frac{\mu^k}{2} \left\| \frac{\mathbf{\Pi}^k}{\mu^k} + \mathbf{X} - \mathbf{XZ}^k - \mathbf{E}^k \right\|_F^2. \end{aligned} \tag{5}$$

2.2. Lat-LRR. LRR has two conditions; one is that the original data $\mathbf{X}$ contains enough samples, and the other is that $\mathbf{X}$ contains enough nonpolluting data. However, these two conditions are almost impossible to achieve for gene data. On the one hand, the available number of gene samples for research is small because of the high prices of gene sequencing. On the other hand, due to process, instrument electromagnetic interference, and other factors, noise pollution will be produced inevitably in the process of genetic sequencing. To overcome the limitation of LRR, [25] proposed a method of Lat-LRR which expressed the original observation data $\mathbf{X}$ as a linear combination of principal feature $\mathbf{XZ}$ and latent feature $\mathbf{LX}$ for feature extraction. Considering the characteristics of heavy noise in gene expression profile, we added sparsity constraints to the model to construct the following Lat-LRR function:

$$O_2 = \begin{cases} \min & \|\mathbf{Z}\|_* + \|\mathbf{L}\|_* + \lambda \|\mathbf{E}\|_{2,1} \\ \text{s.t.} & \mathbf{X} = \mathbf{XZ} + \mathbf{LX} + \mathbf{E}. \end{cases} \tag{6}$$

The solution of Lat-LRR is given in [26]. Alternating direction method (ADM) is adopted to solve the model (6). Let $\mathbf{Z} = \mathbf{J}_1$, $\mathbf{L} = \mathbf{J}_2$; we constructed the following Augmented Lagrangian function:

$$\begin{aligned} \mathscr{L} = {} & \left\|\mathbf{J}_1\right\|_* + \left\|\mathbf{J}_2\right\|_* + \lambda \|\mathbf{E}\|_{2,1} + \left\langle \mathbf{\Lambda}, \mathbf{Z} - \mathbf{J}_1 \right\rangle + \left\langle \mathbf{\Pi}, \mathbf{L} \right. \\ & \left. - \mathbf{J}_2 \right\rangle + \langle \mathbf{\Delta}, \mathbf{X} - \mathbf{XZ} - \mathbf{LX} - \mathbf{E} \rangle + \frac{\mu}{2} \left\{ \left\|\mathbf{Z} - \mathbf{J}_1\right\|_F^2 \right. \\ & \left. + \left\|\mathbf{L} - \mathbf{J}_2\right\|_F^2 + \|\mathbf{X} - \mathbf{XZ} - \mathbf{LX} - \mathbf{E}\|_F^2 \right\}. \end{aligned} \tag{7}$$

Keep $\mathbf{Z} = \mathbf{Z}^k$ and $\mathbf{\Lambda} = \mathbf{\Lambda}^k$; update $\mathbf{J}_1$:

$$\mathbf{J}_1^{k+1} = \arg\min_{\mathbf{J}_1} \left\| \mathbf{J}_1^k \right\|_* + \frac{\mu^k}{2} \left\| \frac{\mathbf{\Lambda}^k}{\mu^k} + \mathbf{Z}^k - \mathbf{J}_1^k \right\|_F^2. \tag{8}$$

Keep $\mathbf{L} = \mathbf{L}^k$, $\mathbf{\Pi} = \mathbf{\Pi}^k$; update $\mathbf{J}_2$:

$$\mathbf{J}_2^{k+1} = \arg\min_{\mathbf{J}} \left\| \mathbf{J}_2^k \right\|_* + \frac{\mu^k}{2} \left\| \frac{\mathbf{\Pi}^k}{\mu^k} + \mathbf{L}^k - \mathbf{J}_2^k \right\|_F^2. \tag{9}$$

Keep $\mathbf{J}_1 = \mathbf{J}_1^k$, $\mathbf{L} = \mathbf{L}^k$, $\mathbf{E} = \mathbf{E}^k$, $\mathbf{\Lambda} = \mathbf{\Lambda}^k$, and $\mathbf{\Delta} = \mathbf{\Delta}^k$; update $\mathbf{Z}$:

$$\begin{aligned} \mathbf{Z}^{k+1} = {} & \left(\mathbf{I} + \mathbf{X}^T \mathbf{X} \right)^{-1} \\ & \cdot \left(\mathbf{X}^T \left(\mathbf{X} - \mathbf{L}^k \mathbf{X} - \mathbf{E}^k \right) + \mathbf{J}_1^k + \frac{\left(\mathbf{X}^T \mathbf{\Delta}^k - \mathbf{\Lambda}^k \right)}{\mu^k} \right). \end{aligned} \tag{10}$$

Keep $\mathbf{Z} = \mathbf{Z}^k$, $\mathbf{E} = \mathbf{E}^k$, $\mathbf{\Pi} = \mathbf{\Pi}^k$, and $\mathbf{\Delta} = \mathbf{\Delta}^k$; update $\mathbf{L}$:

$$\begin{aligned} \mathbf{L}^{k+1} = {} & \left(\left(\mathbf{X} - \mathbf{XZ}^k - \mathbf{E}^k \right) \mathbf{X}^T + \mathbf{J}_2^k + \frac{\left(\mathbf{\Delta}^k \mathbf{X}^T - \mathbf{\Pi}^k \right)}{\mu^k} \right) \\ & \cdot \left(\mathbf{I} + \mathbf{XX}^T \right)^{-1}. \end{aligned} \tag{11}$$

Keep $\mathbf{Z} = \mathbf{Z}^k$, $\mathbf{L} = \mathbf{L}^k$; update $\mathbf{E}$:

$$\begin{aligned} \mathbf{E}^{k+1} = \arg\min_{\mathbf{E}} & \, \lambda \left\| \mathbf{E}^k \right\|_{2,1} \\ & + \frac{\mu^k}{2} \left\| \mathbf{X} - \mathbf{XZ}^k - \mathbf{L}^k \mathbf{X}^k - \mathbf{E}^k \right\|_F^2. \end{aligned} \tag{12}$$

3. Method

3.1. NNDGLLRR. Lat-LRR overcomes the problem of too many constraints of LRR dictionary; however, Lat-LRR has limited ability to recover the subspace, and too many auxiliary variables are involved in the process of algorithm solving that involves a lot of matrix singularity value decomposition (SVD) and matrix inversion, which will affect the performance of the algorithm. Ref. [27] proposed a feature extraction method combining manifold constraint and nonnegative matrix factorization (NMF). In the case of NMF reducing dimensionality, the internal spatial structure of the data is maintained by manifold regularized constraint, and good experimental results are obtained. Ref. [28, 29] proposed an image clustering method combining manifold regularized constraint with Lat-LRR. Similar to the image data, the gene expression profile is also constituted by numerical matrix with high redundancy and heavy noise. Considering this characteristic, we constructed a new NNDGLLRR model on the basis of the original model.

$$O_3 = \begin{cases} \min\limits_{\mathbf{Z}} & \|\mathbf{Z}\|_* + \|\mathbf{L}\|_* + \frac{\alpha}{2}\operatorname{Tr}\left(\mathbf{Z}\mathbf{S}_1\mathbf{Z}^T\right) + \frac{\beta}{2}\operatorname{Tr}\left(\mathbf{L}\mathbf{S}_2\mathbf{L}^T\right) + \lambda\|\mathbf{E}\|_{2,1} \\ \text{s.t.} & \mathbf{X} = \mathbf{L}\mathbf{X}\mathbf{Z} + \mathbf{E}, \quad \mathbf{Z} \geq 0, \ \mathbf{L} \geq 0, \end{cases} \tag{13}$$

where α, β, and λ are nonnegative constants; the model is a nonnegative latent low-rank representation (NNLLRR) when α and β are equal to zero. Model (13) takes a more general form. The dual regularized constraint is used to preserve the internal spatial structure of the original data, and sparse constraints and nonnegative constraints are used to maintain and enhance the robustness of the algorithm. $\mathbf{S}_1$ and $\mathbf{S}_2$ are Laplacian matrices, $\mathbf{S}_1 = \mathbf{D}_1 - \mathbf{W}_1$, $\mathbf{S}_2 = \mathbf{D}_2 - \mathbf{W}_2$. $\mathbf{W}_1$, and $\mathbf{W}_2$ are weight matrix, and there are many ways to solve $\mathbf{W}$, and here we use Gaussian thermal weight. The specific solution is as follows:

$$\begin{aligned} \left(\mathbf{W}_1\right)_{ij} &= e^{-\|x_i - x_j\|_F^2/\sigma}; \quad i, j = 1, 2, 3, \ldots, n \\ \left(\mathbf{W}_2\right)_{ij} &= e^{-\|(x^i)^T - (x^j)^T\|_F^2/\sigma}; \quad i, j = 1, 2, 3, \ldots, m, \end{aligned} \tag{14}$$

where σ is a constant; x_i and x_j represent the ith column and jth column of $\mathbf{X}$ (ith and jth sample); x^i and x^j represent the ith row and the jth row of $\mathbf{X}$, $\mathbf{D}_{ij} = \sum_j \mathbf{W}_{ij}$.

ADM is used to solve model (12), and the following augmented Lagrange function is constructed:

$$\begin{aligned} \mathscr{L} = {} & \|\mathbf{Z}\|_* + \|\mathbf{L}\|_* + \frac{\alpha}{2}\operatorname{Tr}\left(\mathbf{Z}\mathbf{S}_1\mathbf{Z}^T\right) + \frac{\beta}{2}\operatorname{Tr}\left(\mathbf{L}\mathbf{S}_2\mathbf{L}^T\right) \\ & + \lambda\|\mathbf{E}\|_{2,1} + \frac{\mu}{2}\left\|\frac{\mathbf{\Lambda}}{\mu} + \mathbf{X} - \mathbf{L}\mathbf{X}\mathbf{Z} - \mathbf{E}\right\|_F^2, \end{aligned} \tag{15}$$

where $\mathbf{\Lambda}$ is a Lagrangian multiplier; μ is a constant and $\mu > 0$.

Data in real life is generally nonnegative, and nonnegative constraints will make the calculation with a certain degree of sparseness and enhance the robustness of the algorithm. To maintain the nonnegative of feature extraction, we define the following operators:

$$P\left(a_{ij}\right) = \begin{cases} a_{ij}; & \text{if } a_{ij} > 0 \\ 0; & \text{otherwise.} \end{cases} \tag{16}$$

The solution of model (15) is divided into three subproblems: first, the solution of variable $\mathbf{Z}$, second, the solution of variables $\mathbf{L}$, and, third, the solution variable of $\mathbf{E}$.

(1) Solving the First Subproblem. Update $\mathbf{Z}$:

$$\begin{aligned} \mathbf{Z}^{k+1} = {} & \arg\min_{\mathbf{Z}} \ \|\mathbf{Z}\|_* + \frac{\alpha}{2}\operatorname{Tr}\left(\mathbf{Z}\mathbf{S}_1\mathbf{Z}^T\right) \\ & + \frac{\mu^k}{2}\left\|\frac{\mathbf{\Lambda}^k}{\mu^k} + \mathbf{X} - \mathbf{L}^k\mathbf{X}\mathbf{Z} - \mathbf{E}^k\right\|_F^2. \end{aligned} \tag{17}$$

Regarding Taylor second-order expansion to (17), the approximate solution of $\mathbf{Z}$ is as follows:

$$\begin{aligned} \mathbf{Z}^{k+1} = {} & \arg\min_{\mathbf{Z}} \ \|\mathbf{Z}\|_* + \frac{\alpha}{2}\operatorname{Tr}\left(\mathbf{Z}\mathbf{S}_1\mathbf{Z}^T\right) + \frac{\eta_{\mathbf{Z}}\mu^k}{2}\Big\|\mathbf{Z} \\ & - \mathbf{Z}^k - \frac{1}{\eta_{\mathbf{Z}}\mu^k}\mathbf{X}^T\left(\mathbf{L}^k\right)^T\mathbf{\Pi}^k\Big\|_F^2 = \arg\min_{\mathbf{Z}} \ \|\mathbf{Z}\|_* \\ & + \frac{\eta_{\mathbf{Z}}\mu^k + \alpha\|\mathbf{S}_1\|}{2}\Big\|\mathbf{Z} - \mathbf{Z}^k \\ & - \frac{1}{\eta_{\mathbf{Z}}\mu^k + \alpha\|\mathbf{S}_1\|}\left(\mathbf{X}^T\left(\mathbf{L}^k\right)^T\mathbf{\Pi}^k - \alpha\mathbf{Z}^k\mathbf{S}_1\right)\Big\|_F^2 \\ & = \mathbf{D}_{1/\omega_{\mathbf{Z}}}\left(\mathbf{Z}^k + \frac{1}{\omega_{\mathbf{Z}}}\left(\mathbf{X}^T\left(\mathbf{L}^k\right)^T\mathbf{\Pi}^k - \alpha\mathbf{Z}^k\mathbf{S}_1\right)\right). \end{aligned} \tag{18}$$

Nonnegative constraints to $\mathbf{Z}$ are as follows:

$$\left(\mathbf{Z}^{k+1}\right)_{ij} = P\left(\mathbf{Z}^{k+1}\right)_{ij}. \tag{19}$$

Define $\eta_{\mathbf{Z}} = \partial^2 h/\partial\mathbf{Z}^2$; $h = (\mu^k/2)\|\mathbf{\Lambda}^k/\mu^k + \mathbf{X} - \mathbf{L}^k\mathbf{X}\mathbf{Z}^k - \mathbf{E}^k\|_F^2$; $\omega_{\mathbf{Z}} = \eta_{\mathbf{Z}}\mu^k + \alpha\|\mathbf{S}_1\|$; $\mathbf{\Pi}^k = \mathbf{\Lambda}^k + \mu^k(\mathbf{X} - \mathbf{L}^k\mathbf{X}\mathbf{Z}^k - \mathbf{E}^k)$. Ref. [30] gives the solution of $D_\varepsilon(\cdot)$; the solution process is as follows:

$$\begin{aligned} D_\varepsilon(\boldsymbol{\varphi}) &= \mathbf{U}S_\varepsilon(\mathbf{\Omega})\mathbf{V}^T \\ &= \arg\min_{\mathbf{T}} \ \boldsymbol{\varepsilon}\|\mathbf{T}\|_* + \frac{1}{2}\|\mathbf{T} - \boldsymbol{\varphi}\|_F^2. \end{aligned} \tag{20}$$

In (20), $\mathbf{U\Omega V}^T$ is the singular value decomposition (SVD) of $\boldsymbol{\varphi}$, $S_\varepsilon(\cdot)$ is the vector form of the singular value contraction operator (SVT), and $S_\varepsilon(\boldsymbol{\Omega})$ is defined as follows:

$$S_\varepsilon(\boldsymbol{\Omega}) = \operatorname{diag}\left(\operatorname{sgn}\left(\boldsymbol{\Omega}_{\mathbf{ii}}\right)\right)\left(\left|\boldsymbol{\Omega}_{\mathbf{ii}}\right| - \boldsymbol{\varepsilon}\right). \tag{21}$$

(2) Solving the Second Subproblem. Similarly, update $\mathbf{L}$:

$$\begin{aligned}\mathbf{L}^{k+1} &= \arg\min_{\mathbf{L}} \ \|\mathbf{L}\|_* + \frac{\beta}{2}\operatorname{Tr}\left(\mathbf{L}\mathbf{S}_2\mathbf{L}^T\right) + \frac{\eta_{\mathbf{L}}\mu^k}{2}\left\|\mathbf{L} - \mathbf{L}^k \right.\\ &\quad \left. - \boldsymbol{\Pi}^k\left(\mathbf{Z}^k\right)^T\mathbf{X}^T\right\|_F^2 = \arg\min_{\mathbf{L}} \ \|\mathbf{L}\|_* \\ &\quad + \frac{\eta_{\mathbf{L}}\mu^k + \beta\|\mathbf{S}_2\|}{2}\left\|\mathbf{L} - \mathbf{L}^k \right.\\ &\quad \left. - \frac{1}{\eta_{\mathbf{L}}\mu^k + \beta\|\mathbf{S}_2\|}\left(\boldsymbol{\Pi}^k\left(\mathbf{Z}^k\right)^T\mathbf{X}^T - \beta\mathbf{L}^k\mathbf{S}_2\right)\right\|_F^2 \\ &= \mathbf{D}_{1/\omega_{\mathbf{L}}}\left(\mathbf{L}^k + \frac{1}{\omega_{\mathbf{L}}}\left(\boldsymbol{\Pi}^k\left(\mathbf{Z}^k\right)^T\mathbf{X}^T - \beta\mathbf{L}^k\mathbf{S}_2\right)\right).\end{aligned} \tag{22}$$

Nonnegative constraints to $\mathbf{L}$ are as follows:

$$\left(\mathbf{L}^{k+1}\right)_{ij} = P\left(\mathbf{L}^{k+1}\right)_{ij}. \tag{23}$$

Define $\eta_{\mathbf{L}} = \partial^2 h/\partial\mathbf{L}^2$; $\omega_{\mathbf{L}} = \eta_{\mathbf{L}}\mu^k + \beta\|\mathbf{S}_2\|$.

(3) Solving the Third Subproblem. Update $\mathbf{E}$:

$$\begin{aligned}\mathbf{E}^{k+1} &= \arg\min_{\mathbf{E}} \ \lambda\|\mathbf{E}\|_{2,1} + \left\langle \boldsymbol{\Lambda}^k, \mathbf{X} - \mathbf{L}^k\mathbf{X}\mathbf{Z}^k - \mathbf{E}\right\rangle \\ &\qquad + \frac{\mu^k}{2}\left\|\mathbf{X} - \mathbf{L}^k\mathbf{X}\mathbf{Z}^k - \mathbf{E}\right\|_F^2 \\ &= \lambda\|\mathbf{E}\|_{2,1} + \frac{\mu^k}{2}\left\|\frac{\boldsymbol{\Lambda}^k}{\mu^k} + \mathbf{X} - \mathbf{L}^k\mathbf{X}\mathbf{Z}^k - \mathbf{E}\right\|_F^2 \\ &= \boldsymbol{\Theta}_{\lambda/\mu^k}\left(\frac{\boldsymbol{\Lambda}^k}{\mu^k} + \mathbf{X} - \mathbf{L}^k\mathbf{X}\mathbf{Z}^k\right),\end{aligned} \tag{24}$$

where $\Theta_\tau(\cdot)$ is a soft threshold operator (ST); $\Theta_\tau(\cdot)$ is defined as follows:

$$\Theta_\tau(\boldsymbol{\psi}) = \operatorname{sgn}(\boldsymbol{\psi})\max\left(|\boldsymbol{\psi}| - \tau, 0\right). \tag{25}$$

The iterative process of each variable of NNDGLLRR is given above. The concrete updating process is shown in Algorithm 1.

3.2. Sparse Representation Classifier (SRC). Sparse representation is a hotspot in the field of pattern recognition in recent years. SRC has been successfully applied in the field of image classification and has achieved relatively ideal experimental results [31]. Similar to the image data, the gene expression profile is also composed by a series of high redundancy and heavy noise of gene samples. In this paper, the latent features extracted by NNDGLLRR are regarded as data dictionary to construct the following SRC model:

Table 1: Test data information.

Data name	Size	Classes	Number
DLBCL	5469 × 77	2	77
MLL	12582 × 72	3	72
LC	12600 × 203	5	203
ALL	12626 × 248	6	248

$$O_4 = \arg\min_{\boldsymbol{\zeta}} \ \left\|\mathbf{D}\boldsymbol{\zeta} - \mathbf{L}^*\mathbf{y}\right\|_2 + \gamma\|\boldsymbol{\zeta}\|_1. \tag{26}$$

According to the result of SRC, we can get the classification result of unknown gene sample $\mathbf{y}$:

$$i^* = \arg\min_i \ \left\|\mathbf{D}\delta_i(\boldsymbol{\zeta}) - \mathbf{L}^*\mathbf{y}\right\|_2. \tag{27}$$

The detailed flow of the SRC is shown in Algorithm 2.

3.3. Algorithm Flow. To sum up, the algorithm can be divided into two parts; one is to use NNDGLLRR to extract latent features of the original gene expression profile, and the other is to use SRC to classify the latent features. The overall flow is as shown in Algorithm 3.

4. Results and Discussion

4.1. Selecting the Test Data. To test the feature extraction performance of the algorithm, we used diffuse large B-cell lymphoma [32] (DLBCL), mixed lineage leukemia [33] (MLL), lung cancer [34] (LC), acute lymphoblastic leukemia [35] (ALL) gene sequences to make test, and the sample information of each group of genes as is shown in Table 1.

4.2. Accuracy Test. K-means and sparse representation classifier (SRC) are simple and common classifiers. To compare the clustering results of K-means and SRC, the two kinds of classifiers are used to classify the original gene expression profile. Clustering results are shown in Table 2. It is not difficult to find that the classification effect of SRC is significantly higher than that of K-means, which is due to the small number of gene expression profiles. To verify the effectiveness of the algorithm for feature extraction, the extracted features from LRR, Lat-LRR, and NNDGLLRR are classified by SRC. Classification results as shown in Table 2.

Table 2 shows that any one of LRR, Lat-LRR, and NNDGLLRR can achieve feature extraction effectively. However, the feature extraction effect of NNDGLLRR is better than that of Lat-LRR. The category and number of samples, as well as dimension of the gene expression profile, will have an impact on the final recognition effect.

4.3. The Influence of Graph Regularized Coefficients. Generally, we set $\alpha = \beta$. To verify the influence of graph regularized coefficients on feature extraction, we have compared the

```
Input: $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\alpha > 0$, $\beta > 0$, $\gamma > 0$
Initialization: $\mathbf{Z}^0 \in \mathbb{R}^{n\times n}$, $\mathbf{L}^0 \in \mathbb{R}^{m\times m}$, $\mathbf{E}^0 = \mathbf{\Lambda}^0 = \mathbf{\Pi}^0 = \mathbf{0}$, $\mu^0, \mu_{\max}, \rho, \varepsilon_1, \varepsilon_2$
While: $\max(\|\mathbf{Z}^{k+1} - \mathbf{Z}^k\|_\infty, \|\mathbf{L}^{k+1} - \mathbf{L}^k\|_\infty) > \varepsilon_1$ and $\|\mathbf{X} - \mathbf{L}^{k+1}\mathbf{X}\mathbf{Z}^{k+1} - \mathbf{E}^{k+1}\|_\infty > \varepsilon_2$ do
(1) update $\mathbf{\Pi}^k$: $\mathbf{\Pi}^k = \mathbf{\Lambda}^k + \mu^k(\mathbf{X} - \mathbf{L}^k\mathbf{X}\mathbf{Z}^k - \mathbf{E}^k)$
(2) update $\mathbf{Z}^{k+1}$: $(\mathbf{Z}^{k+1})_{ij} = P(\mathbf{Z}^{k+1})_{ij}$
(3) update $\mathbf{L}^{k+1}$: $(\mathbf{L}^{k+1})_{ij} = P(\mathbf{L}^{k+1})_{ij}$
(4) update $\mathbf{E}^{k+1}$: $\mathbf{E}^{k+1} = \mathbf{\Theta}_{\lambda/\mu^k}(\mathbf{\Lambda}^k/\mu^k + \mathbf{X} - \mathbf{L}^k\mathbf{X}\mathbf{Z}^k)$
(5) update $\mathbf{\Lambda}^{k+1}$: $\mathbf{\Lambda}^{k+1} = \mathbf{\Lambda}^k + \mu^k(\mathbf{X} - \mathbf{L}^k\mathbf{X}\mathbf{Z}^k - \mathbf{E}^k)$
(6) update $\mu^{k+1}$: $\mu^{k+1} = \min(\mu_{\max}, \rho\mu^k)$
End while
Output: $\mathbf{L}^*$
```

Algorithm 1: Solving NNDGLLRR model with ALM.

Table 2: Algorithm identification accuracy under different data sets.

Dataset	Accuracy (%)				
	K-means	SRC	LRR + SRC	Lat-LRR + SRC	NNDGLLRR + SRC
DLBCL	46.87	69.79	90.62	89.58	94.79
MLL	50.83	71.43	88.33	90.83	97.50
LC	50.79	73.54	87.83	91.26	98.14
AL	43.33	69.44	83.33	87.22	93.32
Average	47.96	71.05	87.53	89.72	95.94

```
Input: $\mathbf{L}^* \in \mathbf{R}^{m\times m}$
(1) Compute $\mathbf{D} = \mathbf{L}^*\mathbf{X}$
(2) Compute the sparse representation coefficient
      by Eq. (26);
(3) Compute $e_i(y) = \|\mathbf{D}\delta_i(\boldsymbol{\zeta}) - \mathbf{L}^*\mathbf{y}\|_2$;
(4) $i^* = \arg\min_i e_i(y)$;
Output: $i^*$
```

Algorithm 2: The flow of SRC.

```
Input: $\mathbf{X} \in \mathbf{R}^{m\times n}$
(1) Compute $\mathbf{L}^*$ by Eq. (13);
(2) classify the gene expression profile Eq. (27)
Output: $i^*$
```

Algorithm 3: Algorithm flow.

recognition results of LRR, Lat-LRR, and NNDGLLRR under the condition of different α (β) values. The results are shown in Figure 1.

Through the test results of MLL and LC, we can find that manifold regularized constraint has obvious optimization effect on the gene expression profile feature extraction when the values of α and β are appropriate, and it can significantly improve the recognition effect of feature extraction. However, α and β should not be too large or too small. The optimal graph regularized coefficients may be different for different test data sets.

4.4. The Influence of Sparse Representation Coefficients. During the process of gene sequencing, the resulting gene expression profile will usually contain heavy noise due to the sequencing process. To verify the effect of the sparse constraint on the feature extraction, we tested the classification accuracy of LRR, Lat-LRR, and NNDGLLRR for feature extraction under different sparse constraint coefficients λ. The test results are shown in Figure 2.

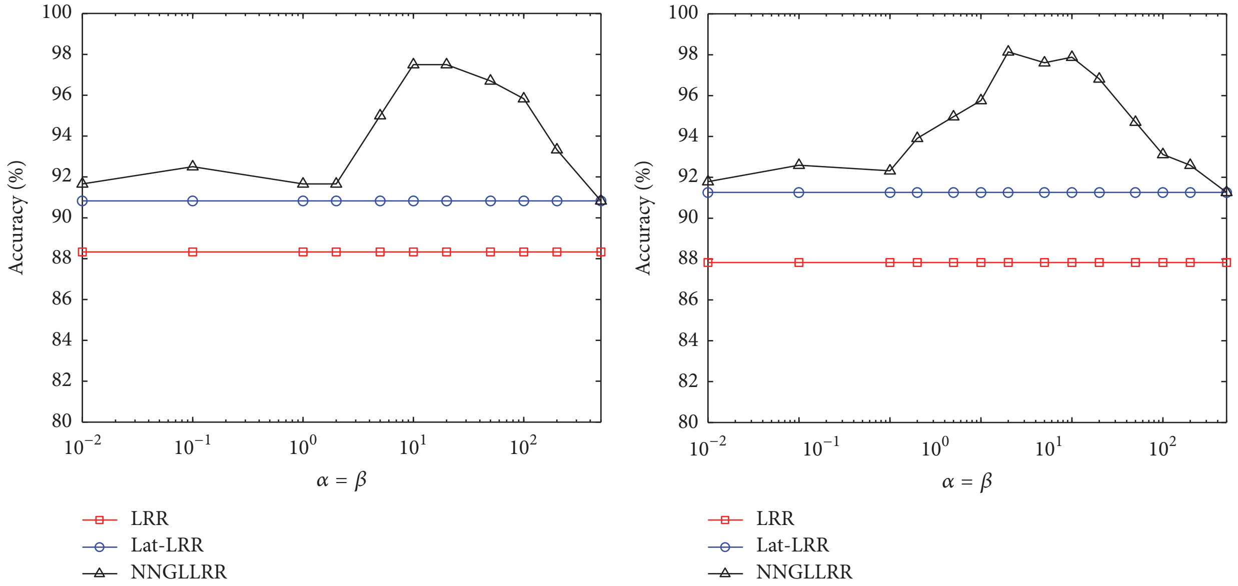
Figure 2 shows that different sparse constraint coefficients have a considerable effect on the final feature extraction results. When the value of λ is appropriate, the performance of Lat-LRR and NNDGLLRR on feature extraction is better than that of LRR. In general, the performance of NNDGLLRR is better than that of Lat-LRR, which proves the validity of manifold constraint again.

4.5. Complexity Analysis. $\mathbf{Z} \in \mathbb{R}^{n\times n}$, $\mathbf{L} \in \mathbb{R}^{m\times n}$, and $\mathbf{E} \in \mathbb{R}^{m\times n}$, and we set the lowest ranks of $\mathbf{Z}$ and $\mathbf{L}$ obtained by the algorithm as r_1 and r_2. Then the complexity of SVT operation for $\mathbf{Z}$ and $\mathbf{L}$ is about $\mathbf{O}(r_1 n^2)$ and $\mathbf{O}(r_2 m^2)$, and the complexity of ST operation for $\mathbf{E}$ is about $\mathbf{O}(kmn)$. The complexity of construction the Laplacian matrix of $\mathbf{Z}$ and $\mathbf{L}$ is about $\mathbf{O}(mn^2)$ and $\mathbf{O}(m^2 n)$; and the complexity of one positive operation for $\mathbf{Z}$ and $\mathbf{L}$ is about n^2 and m^2. If the iteration of the algorithm is k, then the overall complexity of LRR, Lat-LRR, and NNDGLLRR algorithms is shown in Table 3.

Generally, it is considered that $m \gg n$ for gene expression profile. It can be seen from Table 3 that LRR is the simplest in terms of computational complexity, but the performance of LRR on feature extraction is less effective than that of Lat-LRR and NNDGLLRR, and it is difficult to meet the

Table 3: Algorithm complexity calculation.

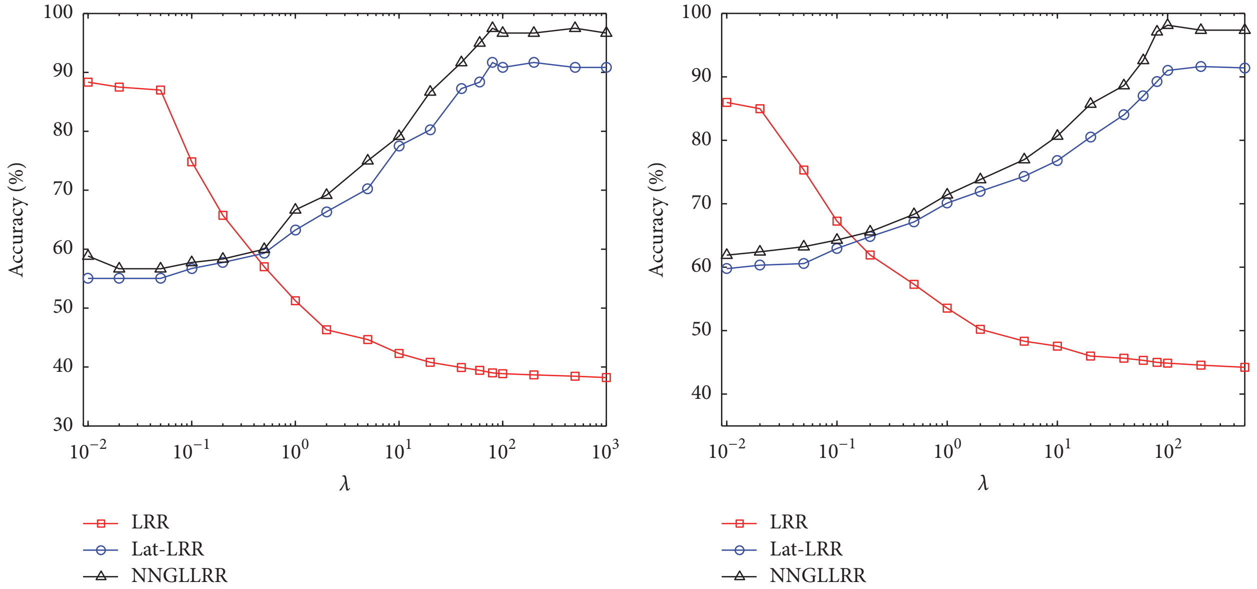
	Complexity			
	SVT	ST	Others	Total
LRR	$2kr_1n^2$	kmn	0	$\mathbf{O}(2kr_1n^2 + kmn)$
Lat-LRR	$2kr_1n^2 + 2kr_2m^2$	kmn	0	$\mathbf{O}(2kr_2m^2)$
NNDGLLRR	$kr_1n^2 + kr_2m^2$	kmn	$mn^2 + m^2n + kn^2 + km^2$	$\mathbf{O}(kr_2m^2)$



(a) Dataset of MLL

(b) Dataset of LC

Figure 1: Clustering performance of algorithms with different graph regularized coefficients.



(a) Dataset of NHL

(b) Dataset of AL

Figure 2: Clustering performance of algorithms with different sparse representation coefficients.

actual demand. The result of Lat-LRR on feature extraction can be not bad, but the partitioning ability of the subspace is limited, and the operation speed is slow because of too many introduced variables. The variable update algorithm of NNDGLLRR not only reduces the calculated amount, but also achieves satisfactory results on feature extraction.

5. Conclusion

Aiming at the characteristics of high redundancy and heavy noise of gene expression profile, a feature extraction model of NNDGLLRR is proposed in this paper. In the process of experiment, we extracted the features of different gene expression profile by LRR, Lat-LRR, and NNDGLLRR and classified the extracted features by SRC. The experimental results show that the performance of NNDGLLRR on feature extraction is better than that of LRR and better than Lat-LRR slightly, which verified the comparative advantages of NNDGLRR. At the same time, compared with Lat-LRR, the overall complexity of NNDGLLRR is reduced through the improvement of the variable update algorithm. The experiments using different gene expression data sets for testing have made comparatively ideal experimental results, which proves the validity of the dual graph regularized constraint. In summary, the proposed nonnegative low-rank sparse constraint and dual graph regularized constraint are reasonable, and NNDGLLRR has good adaptability to different gene expression profile with high redundancy and heavy noise.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper is supported by the National Nature Science Foundation of China (nos. 51365017 and 61305019) and the Science and Technology Project of Jiangxi Province Education Department (no. GJJ150680).

References

[1] S. Bicciato, A. Luchini, and C. Di Bello, "PCA disjoint models for multiclass cancer analysis using gene expression data," *Bioinformatics*, vol. 19, no. 5, pp. 571–578, 2003.

[2] D. Lee, W. Lee, Y. Lee, and Y. Pawitan, "Super-sparse principal component analyses for high-throughput genomic data," *BMC Bioinformatics*, vol. 11, article 296, 2010.

[3] J.-X. Liu, Y.-T. Wang, C.-H. Zheng, W. Sha, J.-X. Mi, and Y. Xu, "Robust PCA based method for discovering differentially expressed genes," *BMC Bioinformatics*, vol. 14, no. 8, article S3, 10 pages, 2013.

[4] A. Sharma and K. K. Paliwal, "Cancer classification by gradient LDA technique using microarray gene expression data," *Data and Knowledge Engineering*, vol. 66, no. 2, pp. 338–347, 2008.

[5] J. Ye, J. Chen, R. Janardan, and S. Kumar, "Developmental stage annotation of Drosophila gene expression pattern images via an entire solution path for LDA," *ACM Transactions on Knowledge Discovery from Data*, vol. 2, no. 1, article 4, 2008.

[6] K. K. Paliwal and A. Sharma, "Improved direct LDA and its application to DNA microarray gene expression data," *Pattern Recognition Letters*, vol. 31, no. 16, pp. 2489–2492, 2010.

[7] D. Lutter, K. Stadlthanner, F. Theis et al., "Analyzing gene expression profiles with ICA," in *Proceedings of the 24th IASTED International Conference on Biomedical Engineering (BioMed '06)*, pp. 25–30, ACTA Press, Innsbruck, Austria, 2006.

[8] I. Hiroyuki, S. Hiroki, A. Kazuhiko et al., "Classification of gastric cancer subtypes by applying ICA to gene expression data and pathway analysis using Bayesian network," *IPSJ SIG Technical Reports*, vol. 2012, no. 12, pp. 1–2, 2012.

[9] H. Chen, Y. Zhang, and I. Gutman, "A kernel-based clustering method for gene selection with gene expression data," *Journal of Biomedical Informatics*, vol. 62, pp. 12–20, 2016.

[10] B. Liu, Q. Cui, T. Jiang, and S. Ma, "A combinational feature selection and ensemble neural network method for classification of gene expression data," *BMC Bioinformatics*, vol. 5, no. 1, pp. 239–242, 2004.

[11] A.-H. Tan and H. Pan, "Predictive neural networks for gene expression data analysis," *Neural Networks*, vol. 18, no. 3, pp. 297–306, 2005.

[12] J. Lee and C. Zhang, "Classification of gene-expression data: the manifold-based metric learning way," *Pattern Recognition*, vol. 39, no. 12, pp. 2450–2463, 2006.

[13] H. Huang and H. Feng, "Gene classification using parameter-free semi-supervised manifold learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 3, pp. 818–827, 2012.

[14] X. Hang, "Cancer classification by sparse representation using microarray gene expression data," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomeidcine Workshops*, pp. 174–177, IEEE, Philadelphia, Pa, USA, November 2008.

[15] C.-H. Zheng, L. Zhang, T.-Y. Ng, C. K. Shiu, and D.-S. Huang, "Metasample-based sparse representation for tumor classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 5, pp. 1273–1282, 2011.

[16] R. Mehra, S. Varambally, L. Ding et al., "Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis," *Cancer Research*, vol. 65, no. 24, pp. 11259–11264, 2005.

[17] G. Ye, M. Tang, J. F. Cai et al., "Correction: low-rank regularization for learning gene expression programs," *PLoS ONE*, vol. 9, no. 1, Article ID e82146, 2014.

[18] A. Kapur, K. Marwah, and G. Alterovitz, "Gene expression prediction using low-rank matrix completion," *BMC Bioinformatics*, vol. 17, no. 1, article 243, 2016.

[19] C.-H. Zheng, T.-Y. Ng, L. Zhang, C.-K. Shiu, and H.-Q. Wang, "Tumor classification based on non-negative matrix factorization using gene expression data," *IEEE Transactions on Nanobioscience*, vol. 10, no. 2, pp. 86–93, 2011.

[20] Y.-S. Lee, A. Krishnan, Q. Zhu, and O. G. Troyanskaya, "Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies," *Bioinformatics*, vol. 29, no. 23, pp. 3036–3044, 2013.

[21] L. Nanni, S. Brahnam, S. Ghidoni, E. Menegatti, and T. Barrier, "A comparison of methods for extracting information from the co-occurrence matrix for subcellular classification," *Expert Systems with Applications*, vol. 40, no. 18, pp. 7457–7467, 2013.

[22] M. Vimaladevi and B. Kalaavathi, "A microarray gene expression data classification using hybrid back propagation neural network," *Genetika*, vol. 46, no. 3, pp. 1013–1026, 2014.

[23] G. Lee, A. Singanamalli, H. Wang et al., "Supervised Multi-view Canonical Correlation Analysis (sMVCCA): integrating histologic and proteomic features for predicting recurrent prostate cancer," *IEEE Transactions on Medical Imaging*, vol. 34, no. 1, pp. 284–297, 2015.

[24] R. Vidal and P. Favaro, "Low rank subspace clustering (LRSC)," *Pattern Recognition Letters*, vol. 43, no. 1, pp. 47–61, 2014.

[25] P. Li, J. Bu, J. Yu, and C. Chen, "Towards robust subspace recovery via sparsity-constrained latent low-rank representation," *Journal of Visual Communication and Image Representation*, vol. 37, pp. 46–52, 2016.

[26] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 1615–1622, November 2011.

[27] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[28] Z. Zhang, S. Yan, and M. Zhao, "Similarity preserving low-rank representation for enhanced data representation and effective subspace learning," *Neural Networks*, vol. 53, pp. 81–94, 2014.

[29] M. Yin, J. Gao, Z. Lin, Q. Shi, and Y. Guo, "Dual graph regularized latent low-rank representation for subspace clustering," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4918–4933, 2015.

[30] Z. Zhang, G. Ely, S. Aeron et al., "Novel methods for multilinear data completion and de-noising based on tensor-SVD," *Computer Science*, vol. 44, no. 9, pp. 3842–3849, 2014.

[31] J. Wang, C. Lu, M. Wang, P. Li, S. Yan, and X. Hu, "Robust face recognition via adaptive sparse representation," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2368–2378, 2014.

[32] M. A. Shipp, K. N. Ross, P. Tamayo et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[33] S. A. Armstrong, J. E. Staunton, L. B. Silverman et al., "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41–47, 2002.

[34] A. Bhattacharjee, W. G. Richards, J. Staunton et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13790–13795, 2001.

[35] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff et al., "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–143, 2002.