Hindawi
Complexity
Volume 2019, Article ID 7081674, 12 pages
https://doi.org/10.1155/2019/7081674

Research Article

Hypergraph Regularized Discriminative Nonnegative Matrix Factorization on Sample Classification and Co-Differentially Expressed Gene Selection

Yong-Jing Hao,[1] **Ying-Lian Gao**,[2] **Mi-Xiao Hou**,[1] **Ling-Yun Dai**,[1] **and Jin-Xing Liu**[1]

[1]*School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China*
[2]*Library of Qufu Normal University, Qufu Normal University, Rizhao 276826, China*

Correspondence should be addressed to Jin-Xing Liu; sdcavell@126.com

Received 29 December 2018; Revised 20 May 2019; Accepted 31 July 2019; Published 19 August 2019

Academic Editor: Mahdi Jalili

Nonnegative Matrix Factorization (NMF) is a significant big data analysis technique. However, standard NMF regularized by simple graph does not have discriminative function, and traditional graph models cannot accurately reflect the problem of multigeometry information between data. To solve the above problem, this paper proposed a new method called Hypergraph Regularized Discriminative Nonnegative Matrix Factorization (HDNMF), which captures intrinsic geometry by constructing hypergraphs rather than simple graphs. The introduction of the hypergraph method allows high-order relationships between samples to be considered, and the introduction of label information enables the method to have discriminative effect. Both the hypergraph Laplace and the discriminative label information are utilized together to learn the projection matrix in the standard method. In addition, we offered a corresponding multiplication update solution for the optimization. Experiments indicate that the method proposed is more effective by comparing with the earlier methods.

1. Introduction

With the development of sequencing technology [1] and gene detection technology [2], a lot of genomic data have been collected. Genomic data usually have the characteristics of high-dimensional small samples, and how to extract useful information from massive genomic data has become the most challenging task. To increase the processing efficiency of such high-dimensional data, a series of dimensionality reduction techniques [3] have been proposed. Among the various dimensionality reduction methods, NMF and its improved NMF-based methods are widely used in the field of gene data processing.

There is some physiology and psychological evidence that humans rely on part-based representations for some object recognition [4]. The Nonnegative Matrix Factorization (NMF) [5] method is capable of learning the various parts of the face and the semantic features of the text. NMF is a powerful technology for component-based data analysis. It is intended to find two nonnegative matrices to learn the part-based representation of the standard data itself. NMF has been popular for decades and successfully implemented in a wide range of fields, including robotics control [6], image analysis [7], and biomedical engineering [8]. To this end, we will provide a brief introduction to the relevant methods.

A variant of NMF called the Graph Regularized Nonnegative Matrix Factorization (GNMF) method [9] was proposed by Cai et al. which takes into consideration the geometric structure between data and uses K-nearest neighbor graph coding to determine the geometry of the data. The method works well in cluster applications but achieves mediocre results in classification problems. To improve the effect of GNMF in the classification, a method named Graph Regularized Discriminative Nonnegative Matrix Factorization (GDNMF) [10] was proposed by Long et al. which considers geometry between data and label information. The discriminating power of different classes is increased by considering the label information. In the GDNMF method, the introduction of the dictionary matrix [11] has achieved excellent performance in the feature selection and classification of

genomic data. The GNMF and GDNMF methods construct a simple graph based on the geometric relationships between the sample data, and the high-order relationships between the sample data may be ignored. So Zeng et al. perfectly integrated the hypergraph regularization into the standard NMF, called Hypergraph Regularized Nonnegative Matrix Factorization (HNMF) [12]. The hypergraph regularized takes into account the intrinsic manifold structure of the sample data. This method encodes the geometric information of the data space by constructing a hypergraph rather than a simple graph. The data representations discovered by HNMF are not only partial but also sparse. So HNMF can show better performance. In addition, Peng et al. proposed an NMF deformation method called Parallel Vector Field Regularized Nonnegative Matrix Factorization for Image Representation [13]. This method can effectively improve the calculation speed. To increase the adaptability of the method and avoid the ambiguity of artificial selection, the Flexible Nonnegative Matrix Factorization with Adaptively Learned Graph Regularization was proposed by Peng et al. [14], which can effectively solve the above problems. In summary, various variants of NMF have their own unique advantages in feature selection or cluster classification.

On the one hand, discriminative Nonnegative Matrix Factorization has also been well applied in other areas, such as image representation [15], image classification [16], and diesel engine fault diagnosis [17]. On the other hand, hypergraph regularization has become more and more popular in recent years, for example, in image click prediction [18], image ranking [19], and image restoration [20].

Inspired by the above work, we propose a novel method called Hypergraph Regularized Discriminative Nonnegative Matrix Factorization (HDNMF), which takes into consideration the sample data with manifold inherent structure. The geometry of the data space is encoded by constructing a hypergraph rather than a simple graph. To consider the part of the data itself and the space characteristics, the label information is seen as a significant factor. We construct the K-nearest neighbor graph [21] to encode the geometry of the data space and increase the discriminative power of different classes. In this paper, an optimization scheme is shown in detail, and the objective function is solved by multiplication iterative update. Experiments demonstrate that our proposed method achieves better results than some NMF variants.

2. Materials and Methods

Let the input matrix $\mathbf{X}$ be data of m rows and n columns, where rows represent genes and columns represent samples. Usually the value of m is large, which may make data processing of poor accuracy. Therefore, the dimensionality reduction of the data matrix has grown up to be a crucial step. The standard NMF and its improved methods are popular techniques for dimensionality reduction. We will introduce the standard NMF method and several standard NMF improvement methods in this section.

2.1. Related Work

2.1.1. Standard NMF (NMF). Nonnegative factorization was provided by Melvyn W. Jeter and Wallace C. Pye [22] in 1984 and the positive matrix factorization by Paatero and Tapper [23] in 1995. Lee and Seung [24, 25] continued to research on it. It was a matrix factorization method that mainly analyzes high-dimensional data matrices with nonnegative factors. Assuming there are three nonnegative matrices $\mathbf{X} \in \mathbf{R}^{m\times n}$, $\mathbf{W} \in \mathbf{R}^{m\times k}$, and $\mathbf{H} \in \mathbf{R}^{k\times n}$, standard NMF can make the following formula established: $\mathbf{X} \approx \mathbf{WH}$.

The objective function of the standard NMF was used to minimize the Euclidean distance [26] between $\mathbf{X}$ and $\mathbf{WH}$ by using the multiplicative update rules:

$$O_1 = \|\mathbf{X} - \mathbf{WH}\|_F^2, \tag{1}$$

where $\|\cdot\|_F$ is the F-norm of the matrix. $\mathbf{X}$ is named the input data matrix, $\mathbf{W}$ is named the basic matrix, and $\mathbf{H}$ is called the coefficient matrix. The elements involved are nonnegative. The update rules are shown as follows in Lee and Seung's paper:

$$\begin{aligned} w_{mk} &\longleftarrow w_{mk}\frac{\left(\mathbf{WH}^T\right)_{mk}}{\left(\mathbf{WHH}^T\right)_{mk}}, \\ h_{kn} &\longleftarrow h_{kn}\frac{\left(\mathbf{W}^T\mathbf{X}\right)_{kn}}{\left(\mathbf{W}^T\mathbf{WH}\right)_{kn}}. \end{aligned} \tag{2}$$

2.1.2. Graph Regularized NMF (GNMF). The GNMF [9] method was used to construct the geometry structure by using the nearest neighbor graph. It minimizes the objective function as follows:

$$O = \|\mathbf{X} - \mathbf{WH}\|_F^2 + \alpha \mathrm{Tr}\left(\mathbf{HLH}^T\right), \tag{3}$$

where $\mathrm{Tr}(\cdot)$ is the trace of the matrix, α is the regularization parameter which controls the smoothness of the new representation, $\mathbf{L}$ is the graph Laplacian matrix ($\mathbf{L} = \mathbf{D} - \mathbf{C}$), $\mathbf{C}$ is the weight matrix of the nearest neighbor graph, and $\mathbf{D}$ is the diagonal matrix.

2.1.3. Graph Regularized Discriminative NMF (GDNMF). The method GDNMF is proposed, which applies the intrinsic simple geometry between data and discriminative label information to design the objective function. The formula is as follows:

$$O = \|\mathbf{X} - \mathbf{WH}\|_F^2 + \alpha \mathrm{Tr}\left(\mathbf{HLH}^T\right) + \beta \|\mathbf{S} - \mathbf{AH}\|_F^2, \tag{4}$$

where $\mathbf{L}$ is the graph Laplacian matrix ($\mathbf{L} = \mathbf{D} - \mathbf{C}$), $\mathbf{C}$ is the weight matrix of the nearest neighbor graph, and $\mathbf{D}$ is the diagonal matrix. $\mathbf{A}$ is initialized randomly in this method. $\mathbf{S}$, $\mathbf{A}$, and $\mathbf{H}$ are nonnegative matrices; α and β are nonnegative regularization parameters.

2.1.4. Hypergraph Regularized NMF (HNMF). The Laplacian eigenmaps (LE) method is a classical manifold method based on simple graphs. The method mainly considers the relationship between two vertices, while the hypergraph considers the relationship between three and more vertices. In the HNMF

method, the hypergraph and NMF combine very well; its objective function is below:

$$O = \|\mathbf{X} - \mathbf{WH}\|_F^2 + \alpha \mathrm{Tr}\left(\mathbf{HL}_{hyper}\mathbf{H}^T\right), \quad (5)$$

where Tr(.) is the trace of the matrix and α is the regularization parameter. $\mathbf{L}_{hyper}$ is the hypergraph Laplacian matrix ($\mathbf{L}_{hyper} = \mathbf{D}_{hyper} - \mathbf{C}_{hyper}$), $\mathbf{C}_{hyper}$ is the weight matrix of the nearest neighbor graph, and $\mathbf{D}_{hyper}$ is the diagonal matrix.

2.2. Methodology. Inspired by the GDNMF method and HNMF method, we added the discriminative label information to the HNMF method. The definition and multiplication update rules for HDNMF are shown below.

2.2.1. Hypergraph Regularized Discriminative NMF (HDNMF). In a simple graph, two vertices are connected by an edge, and the weight of the edge is utilized to represent the affinity relationship between the two vertices. In fact, the interrelationship between multiple vertices is also essential. To solve this problem, the hypergraph [27, 28] emerges, and its edges can establish a link between two or more vertices.

The hypergraph G_{hyper} consists of V_{hyper}, E_{hyper}, and $\mathbf{W}_{hyper}$, where V_{hyper} represents the set of vertices; E_{hyper} represents the hyperedge set; $\mathbf{W}_{hyper}$ represents the weight set of the hyperedge; the weight of the hyperedge e is represented as $w_{hyper}(e)$. The correlation matrix $\mathbf{H}_{hyper}$ of hypergraph G_{hyper} is given below [28]:

$$\mathbf{H}_{hyper}(v, e) \begin{cases} 1, & if\ v \in e, \\ 0, & if\ v \in e. \end{cases} \quad (6)$$

The degree of one edge e is as follows:

$$d(v) = \sum_{\{e \in E | v \in e\}} w_{hyper}(e) = \sum_{e \in E} w_{hyper}(e)\, \mathbf{H}_{hyper}(v, e). \quad (7)$$

The degree of one hyperedge e is as follows:

$$d_{hyper}(e) = |e| = \sum_{v \in V} \mathbf{H}_{hyper}(v, e). \quad (8)$$

Let $\mathbf{D}_{hyper}$ be a diagonal matrix, which corresponds to the element corresponding to the vertex degree; then the nonstandardized superlattices matrix is as follows:

$$\mathbf{L}_{hyper} = \mathbf{D}_{hyper} - \mathbf{C}_{hyper}, \quad (9)$$

where $\mathbf{C}_{hyper} = \mathbf{H}_{hyper}\mathbf{W}_{hyper}\mathbf{D}_{hyper_e}^{-1}\mathbf{H}_{hyper}^T$.

A label matrix $\mathbf{S}$ is defined as follows:

$$\mathbf{s}_{i,j} = \begin{cases} 1, & if\ y_i = i,\ j = 1, 2, 3 \ldots, n,\ i = 1, 2, 3 \ldots, c, \\ 0, & otherwise, \end{cases} \quad (10)$$

where $y_j \in \{1, 2, 3 \ldots, c\}$ represents the jth sample of the class label of x_j, and the number of categories in the training set $\mathbf{X}$ is c.

We can obtain the following minimization problem:

$$O = \|\mathbf{X} - \mathbf{WH}\|_F^2 + \alpha \mathrm{Tr}\left(\mathbf{HL}_{hyper}\mathbf{H}^T\right) + \beta \|\mathbf{S} - \mathbf{AH}\|_F^2, \quad (11)$$

where Tr(.) is the trace of the matrix, $\mathbf{L}_{hyper}$ is the hypergraph Laplacian matrix ($\mathbf{L}_{hyper} = \mathbf{D}_{hyper} - \mathbf{C}_{hyper}$), $\mathbf{C}_{hyper}$ is the weight matrix of the nearest neighbor graph, and $\mathbf{D}_{hyper}$ is the diagonal matrix. $\mathbf{A}$ is initialized randomly in this method. $\mathbf{S}$, $\mathbf{A}$, and $\mathbf{H}$ are nonnegative matrices; α and β are nonnegative regularization parameters.

2.2.2. The Update Rules of HDNMF. The multiplication update rules are extended according to the standard NMF's F-norm [29], to find the local optimum. Equation (11) can be written as follows:

$$\begin{aligned} F &= \mathrm{Tr}\left((\mathbf{X} - \mathbf{WH})^T (\mathbf{X} - \mathbf{WH})\right) + \alpha \mathrm{Tr}\left(\mathbf{HL}_{Hyper}\mathbf{H}^T\right) \\ &\quad + \beta \mathrm{Tr}\left((\mathbf{S} - \mathbf{AH})^T (\mathbf{S} - \mathbf{AH})\right) \\ &= \mathrm{Tr}\left(\mathbf{X}^T\mathbf{X}\right) - 2\mathrm{Tr}\left(\mathbf{X}^T\mathbf{WH}\right) + \mathrm{Tr}\left(\mathbf{H}^T\mathbf{W}^T\mathbf{WH}\right) \\ &\quad + \beta \mathrm{Tr}\left(\mathbf{S}^T\mathbf{S}\right) - 2\beta \mathrm{Tr}\left(\mathbf{S}^T\mathbf{AH}\right) + \beta \mathrm{Tr}\left(\mathbf{H}^T\mathbf{A}^T\mathbf{AH}\right) \\ &\quad + \alpha \mathrm{Tr}\left(\mathbf{HL}_{Hyper}\mathbf{H}^T\right). \end{aligned} \quad (12)$$

Equation (12) can be written as a Lagrange function as follows:

$$\begin{aligned} L_f &= \mathrm{Tr}\left(\mathbf{X}^T\mathbf{X}\right) - 2\mathrm{Tr}\left(X^T\mathbf{WH}\right) + \mathrm{Tr}\left(\mathbf{H}^T\mathbf{W}^T\mathbf{WH}\right) \\ &\quad + \beta \mathrm{Tr}\left(\mathbf{S}^T\mathbf{S}\right) - 2\beta \mathrm{Tr}\left(\mathbf{S}^T\mathbf{AH}\right) \\ &\quad + \beta \mathrm{Tr}\left(\mathbf{H}^T\mathbf{A}^T\mathbf{AH}\right) + \alpha \mathrm{Tr}\left(\mathbf{HL}_{Hyper}\mathbf{H}^T\right) \\ &\quad + \mathrm{Tr}\left(\Phi^T\mathbf{W}\right) + \mathrm{Tr}\left(\Psi^T\mathbf{H}\right) + \mathrm{Tr}\left(\Omega^T\mathbf{A}\right), \end{aligned} \quad (13)$$

where Φ, Ψ, and Ω are the Lagrange multipliers.

The partial derivatives of L_f with respect to $\mathbf{H}$, $\mathbf{W}$, and $\mathbf{A}$, respectively, are

$$\begin{aligned} \frac{\partial L_f}{\partial \mathbf{H}} &= 2\mathbf{W}^T(\mathbf{WH} - \mathbf{X}) + 2\alpha \mathbf{HL}_{Hyper} \\ &\quad + 2\beta \mathbf{A}^T(\mathbf{AH} - \mathbf{S}) + \Psi = 0. \end{aligned} \quad (14)$$

$$\frac{\partial L_f}{\partial \mathbf{W}} = 2\mathbf{WHH}^T - 2\mathbf{XH}^T + \Phi = 0. \quad (15)$$

$$\frac{\partial L_f}{\partial \mathbf{A}} = -2\beta \mathbf{SH}^T + 2\beta \mathbf{AHH}^T + \Omega = 0. \quad (16)$$

The following formulas can be obtained by using the KKT condition:

$$\begin{aligned} &\left[2\mathbf{W}^T(\mathbf{WH} - \mathbf{X}) + 2\alpha \mathbf{H}\left(\mathbf{D}_{hyper} - \mathbf{C}_{hyper}\right)\right. \\ &\quad \left. + 2\beta \mathbf{A}^T(\mathbf{AH} - \mathbf{S})\right]_{kn} \mathbf{H}_{kn} + \Psi_{kn}\mathbf{H}_{kn} = 0. \end{aligned} \quad (17)$$

$$\left(2\mathbf{WHH}^T - 2\mathbf{XH}^T\right)_{ik} \mathbf{W}_{ik} + \Phi_{ik}\mathbf{W}_{ik} = 0. \quad (18)$$

$$\left(-2\beta \mathbf{SH}^T + 2\beta \mathbf{AHH}^T\right)_{pk} \mathbf{A}_{pk} + \Omega_{pk}\mathbf{A}_{pk} = 0. \quad (19)$$

Input: Data matrix $\mathbf{X} \in \mathbf{R}^{m\times n}$, hyper-graph Laplace matrix $\mathbf{L}_{hyper} \in \mathbf{R}^{n\times n}$, indicator matrix $\mathbf{S} \in \mathbf{R}^{p\times k}$, parameters α, β, k.
Initialization: Randomly initialize three non-negative matrices $\mathbf{W} \in \mathbf{R}^{m\times k}$ $\mathbf{H} \in \mathbf{R}^{k\times n}$ and $\mathbf{A} \in \mathbf{R}^{p\times k}$.
Repeat:
(1) Update **H** by rule (20).
(2) Update **W** by rule (21).
(3) Update **A** by rule (22).
Until Convergence.
Output: **W** and **H**.

Algorithm 1: The algorithm of HDNMF.

Table 1: Computational operation counts for each iteration in NMF and HDNMF.

Algorithm	NMF	HDNMF
Addition	$2mnk + 2(m+n)k^2$	$2mnk + (2m+4n+2c)k^2 + (2n+Kn+cp)k$
Multiplication	$2mnk + 2(m+n)k^2 + (m+n)k$	$2mnk + (2m+4n+2c)k^2 + (3n+Kn+m+cp+c)k$
Division	$(m+n)k$	$(m+n+c)k$
Overall	$O(mnk)$	$O(mnk)$

Table 2: Parameters description.

Parameters	Description
m	Number of features
n	Number of samples
k	Number of dimension
c	Number of class
p	Number of training samples obtained from each class
K	Number of nearest neighbors

We can gain the updating rules for **H**, **W,** and **A**:

$$\mathbf{H}_{kn} \longleftarrow \mathbf{H}_{kn}\frac{\left(\beta\mathbf{A}^T\mathbf{S} + \mathbf{W}^T\mathbf{X} + \alpha\mathbf{H}\mathbf{C}_{hyper}\right)_{kn}}{\left(\mathbf{W}^T\mathbf{W}\mathbf{H} + \beta\mathbf{A}^T\mathbf{A}\mathbf{H} + \alpha\mathbf{H}\mathbf{D}_{hyper}\right)_{kn}}. \quad (20)$$

$$\mathbf{W}_{mk} \longleftarrow \mathbf{W}_{mk}\frac{\left(\mathbf{X}\mathbf{H}^T\right)_{mk}}{\left(\mathbf{W}\mathbf{H}\mathbf{H}^T\right)_{mk}}. \quad (21)$$

$$\mathbf{A}_{pk} \longleftarrow \mathbf{A}_{pk}\frac{\left(\mathbf{S}\mathbf{H}^T\right)_{pk}}{\left(\mathbf{A}\mathbf{H}\mathbf{H}^T\right)_{pk}}. \quad (22)$$

The algorithm of HDNMF is shown in Algorithm 1.

2.2.3. Complexity Analysis. In this subsection, we compare the computational complexity of NMF and HDNMF based on multiplication of update rules. The calculation operation count of every iteration is considered by (2) and (20)-(22). The results and parameters are listed in Tables 1 and 2.

For the HDNMF method, the matrix **C** is sparse. So, we need Knk multiplication and addition to count **HC**. As for the label matrix **S**, we need nk multiplication and addition to count $\mathbf{A}^T\mathbf{S}$, and we need cpk multiplication and addition to count $\mathbf{S}\mathbf{H}^T$.

In addition to the multiplication update, HDNMF requests $O(p^2m)$ to set up the weight matrix **C** and requests $O(pc)$ to set up the indicator matrix **S**. Assume that, after the multiplication iteration of t multiplication, the total cost of NMF is $O(tmnk)$; the total of HDNMF the cost is $O(tmnk + p^2m + pc)$.

3. Results and Discussion

On the one hand, HDNMF considers the high-order relationship about samples; on the other hand, the HDNMF method uses the label information to make the method discriminative while constructing the internal geometry of the data. To evaluate the effectiveness and discrimination of the method, the HDNMF method was compared with the other methods (NMF, DNMF, LNMF, GNMF, GDNMF, and HNMF).

3.1. Datasets Description. The Cancer Genome Atlas (TCGA), as the largest cancer genome database, has immeasurable information value. The data included cholangiocarcinoma data (CHOL), esophageal cancer data (ESCA), pancreatic cancer data (PAAD), colorectal cancer data (COAD), and head and neck squamous cell carcinoma (HNSC) data. All of this data can be obtained from the TCGA database at https://cancergenome.nih.gov/. Each dataset consists of two classes, normal samples and diseased samples. The dimension of the datasets is 20502. The rows of the data represent gene features, and the columns represent gene samples.

Firstly, the CHOL data contained 9 normal samples and 36 diseased samples, the PAAD data contained 4 normal samples of 176 diseased samples, the HNSC data contained 20 normal samples and 198 diseased samples, the ESCA data contained 9 normal samples and 183 diseased samples, and the COAD data contained 19 normal samples and 262 diseased samples. Then, the normal samples are removed. Finally, we integrate the PAAD, ESCA, and CHOL data into datasets with dimensions of 20502∗395 (INTA 1) and

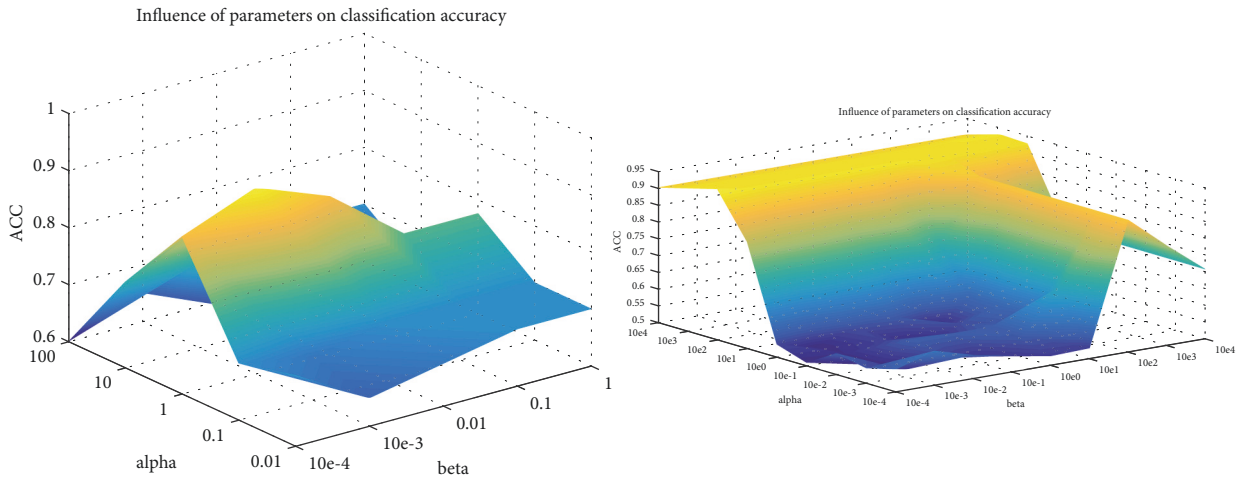

Figure 1: The influence of parameters on classification ACC (INTA 1 and INTA 2).

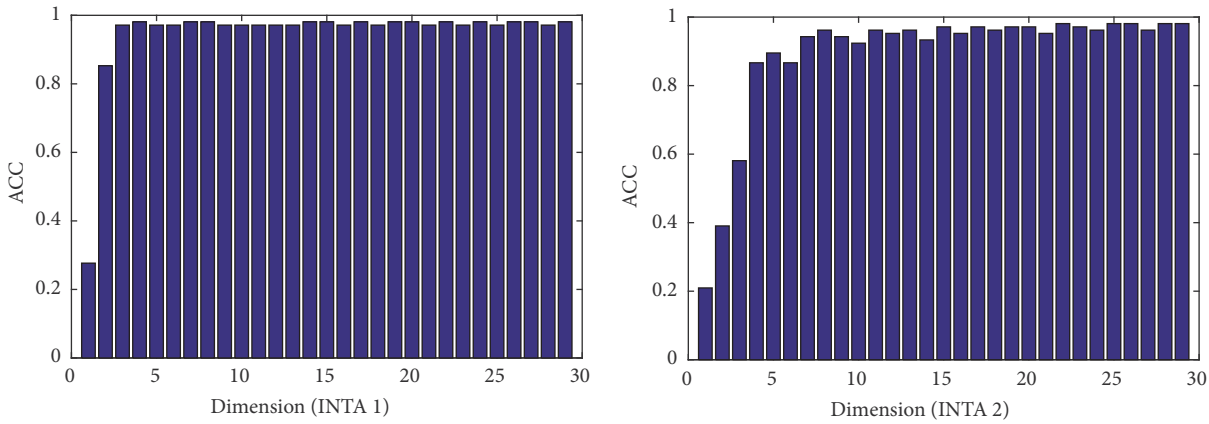

Figure 2: Influence of dimension on classification ACC.

integrate the PAAD, COAD, HNSC, ESCA, and CHOL data into one dataset with dimensions of 20502∗1055 (INTA 2).

3.2. Implementation Issues

3.2.1. Parameter Selection. As for the experiments, the datasets used were INTA 1 and INTA 2. For each test, 10-fold cross validation uses training data to adjust parameters. The advantage of cross validation is that all test datasets are independent. The regularized parameter matrix directly affects the classification results of the HDNMF method. In order to select the optimal parameters, we adjust the regularized parameters exponentially in a specific domain, such as from 10^{-4} to 10^{4}. This section uses the tenfold cross validation method to automatically adjust the selection of the best regularized parameters in the corresponding training datasets in the following $\{10^{t} : t \in \{-4, -3, \ldots 3, 4\}\}$ range. The results were shown in Figure 1.

As shown in the figure, the warmer the color, the higher the classification ACC. As can be seen from Figure 1, dataset INTA 1 can get better results when $\alpha = 1$ and $\beta \in [10^{-4}, 10^{-2}]$; as for dataset INTA 2, it can get better results when $\alpha \in [10^{2}, 10^{4}]$ and $\beta \in [10^{-4}, 10^{4}]$. In particular, the value of α has small significant effect on the classification result when $\beta = 10^{2}$.

3.2.2. The Influence of Dimensions on Classification ACC. In this part, we do the HDNMF method when dimension is from 1 to 30 to explore the effect of dimension size on the classification effect. Based on the results in Figure 2, the following points can be summarized:

(1) When the dimension is too small, the classification ACC effect of all methods is not ideal, mainly due to the loss of a large amount of useful information due to the large dimensionality reduction

(2) As the dimension increases, the effect of the classification ACC gradually increases and tends to be stable after a certain degree, which is mainly due to the large amount of information loss that can be avoided when the dimension is large enough

3.2.3. HDNMF Time Comparison. We compare the average runtime of all the methods on the two datasets in Table 4. The experimental results were obtained by performing all methods on an Inter(R) Core(TM) i7-7700 CPU @ 3.6 GHz

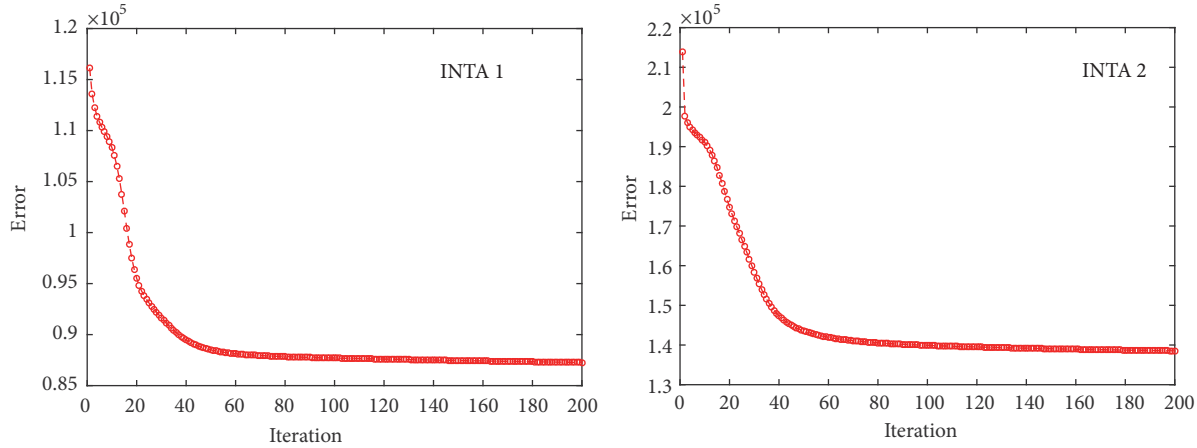

Figure 3: The convergence performance for INTA1 and INTA2.

Table 3: The classification ACC for all methods.

Methods	ACC (INTA 1)	ACC (INTA 2)
NMF	80.00±0.00181	87.62±0.00195
DNMF	85.10±0.00390	90.79±0.00137
LNMF	83.77±0.00690	82.07±0.004504
GNMF	82.50±0.00140	89.52±0.00135
GDNMF	90.00±0.00009	90.48±0.00019
HNMF	87.50±0.00198	88.57±0.00175
HDNMF	*92.50±0.00096*	*94.88±0.00082*

Windows Server with 64 G RAM. The time costs of the classification process for the all methods are listed. Obviously, the NMF method with constraints usually requires more runtime, and our proposed HDNMF method requires the longest runtime. The main reason is that the HDNMF method takes a lot of time in constructing the hypergraph and updates the hyperedge in each iterate. Therefore, how to speed up the proposed algorithm is an interesting work.

3.2.4. The Convergence Analysis. The updating rules for the HDNMF method in (20)-(22) are easy to understand and it can be demonstrated that the above rules are convergent. Figure 3 shows the convergence about two datasets; the red dashed lines are the error value of the HDNMF method. We can see that our method will converge in 100 iterations.

3.3. Classification. Since HDNMF introduces label information into the NMF method, it is necessary to perform the test of the effect of the classification [28]. There are many improved methods for existing NMFs for classification, such as methods which were mentioned in [30, 31].

Classification is a part of the most widely used technologies in the field of data mining. It is a technique for constructing a classifier based on the characteristics of experimental data and then using the classifier to classify samples of unknown categories. The classifier is generally divided into two stages of training and testing; in the training stage, the characteristics of the training datasets are first analyzed, and then an accurate description of the corresponding datasets is generated for each category; in the testing phase, the model classifies the test set to test its classification ACC. In the experimental part, we use the KNN method to classify the sample. For a given set of training samples with class labels, the similarity measure function is used to find the nearest neighbors that are most similar to the samples to be classified, and the decision is made according to the corresponding principle.

The specific classification process (K-nearest neighbors, KNN) is as follows: (1) for the training sets, calculate the distance from the unknown sample to all known samples; (2) select the parameter K (where K is the most basic parameter in KNN, representing any number of neighbors); (3) for test sets, the unknown samples are classified into the largest number of categories in the sample according to the majority voting rule.

3.3.1. HDNMF for INTA 1 and INTA 2 Classification. For the dimension equal to 3 for INTA 1 and 5 for INTA 2 the classification ACC are listed in Table 3. The optimal results are italic, and the experiment proves that our HDNMF method can achieve better classification results than other methods.

It is possible to conclude that the introduction of label information is the main reason for the significant classification ACC effect. Therefore, the HDNMF method can not only improve the interpretability of the NMF method, but also overcome the ambiguity of the unsupervised learning training samples.

Table 4: The time comparison for all methods.

Methods	Time/S (INTA 1)	Time/S (INTA 2)
NMF	6.4064	14.7599
DNMF	7.9114	22.3321
LNMF	7.4628	19.1296
GNMF	7.8372	21.0590
GDNMF	5.0958	16.3989
HNMF	56.5247	93.0869
HDNMF	*29.3517*	*68.6402*

Table 5: NUM, TRS, and ARS of five methods (INTA 1).

Methods		HDNMF	NMF	DNMF	LNMF	GNMF	HNMF	GDNMF
PAAD	Num	*331*	327	324	328	327	322	326
	TRS	*2385.79*	2228.45	2351.36	2287.7	2339.88	2061.11	2350.54
	ARS	*7.2078*	6.8148	7.2350	6.9535	7.1556	6.4010	7.2102
ESCA	Num	252	*257*	248	253	254	243	248
	TRS	*2318.07*	2305.29	2194.42	2231.10	2224.08	2185.63	2194.08
	ARS	*9.1987*	8.9700	8.8485	8.8186	8.7562	8.9944	8.8471
CHOL	Num	*134*	132	134	130	132	137	131
	TRS	*546.15*	531.06	527.15	525.56	526.42	504.76	526.63
	SRS	*4.0758*	4.0232	3.9936	4.0428	3.9880	3.6844	4.0201

Table 6: Co-differentially expressed genes of five methods (INTA 1).

Methods	Codifferentially expressed genes	Unique
NMF	MUC6 FSCN1 POSTN HIF1A SLC7A5 S100A2 *TKT* TFRC TTR *TP63* SPINK1	2
DNMF	SPINK1CTGFS100A2ANPEP	0
LNMF	CRPHIF1AEGR1STAT3SLC7A5TFRC	0
GNMF	MUC6 FSCN1 POSTN SLC7A5 S100A2 TTR SPINK1 TPM1 CTGF	0
HNMF	MUC6 FSCN1 POSTN SLC7A5 TTR TPM1 HIF1A TFRC CRP STAT3 EGR1	0
HDNMF	*ABCC3 BCL2L1* CRP *RHOC* TPM1 *ERBB3 AGT CTSC HP* ANPEP	7

3.4. Co-Differentially Expressed Gene Selection. High-dimensional small sample data contain many unrelated or redundant features. In biological processes, most genes serve as the standard function of life support, so researchers focus on a small number of differentially expressed genes that play a critical role in life. Feature selection [32] is the easiest way to resolve these problems and identify differentially expressed genes. It can replace the standard data by choosing the largest amount of information without losing a lot of information. Here we show the experimental results and analysis of different methods for selecting common differentially expressed genes (co-differentially expressed genes) on integrated datasets. The relevant score (RS) refers to the correlation between genes and diseases. The higher the relevant score, the higher the correlation between genes and diseases. The number of genes (NUM) refers to the number of co-differentially expressed genes selected by each method characteristic that match the proven disease genes in GeneCards.

How the GeneCards were used is described as follows: we visited the official website of GeneCards (https://www.genecards.org/). On the one hand, all types of diseased genes that have been validated can be obtained by entering the type of cancer at this web address (exporting the table for the determination of NUM). On the other hand, enter the gene name at this web address to get the details of the gene.

3.4.1. INTA 1 Co-Differentially Expressed Gene Selection Results. The HDNMF method and the NMF series of methods are used to select co-differentially expressed genes of the integrated datasets. We selected 500 genes for each method to evaluate the co-differentially expressed genes selected by using different methods.

For the 500 co-differentially expressed genes selected by the seven methods, we compare the three diseases that make up the integrated INTA 1. The comparison data are from GeneCards (https://www.genecards.org/). The co-differentially expressed genes excavated by different methods are listed in Table 5: the number of genes (NUM), the total related score (TRS), and the average related score (ARS). Higher NUM, TRS, and ARS values indicate better method performance. From the table, we can see that the HDNMF method can achieve better experimental results.

We summarize the co-differentially expressed genes in the shared part in Table 6, including the co-differentially

Table 7: Unique co-differentially expressed genes by HDNMF (INTA 1).

Official gene symbol	Official name	Related diseases	Related GO annotations	RS	Paralog gene
ERBB3	Erb-B2 receptor tyrosine kinase 3	Retroperitoneal leiomyosarcoma and fatal congenital contracture syndrome 2	Protein homodimerization activity and transferase activity, transferring phosphorus-containing groups	94.06	ERBB4
CTSC	Cathepsin C	Papillon-Lefevre syndrome and Haim-Munk syndrome	Identical protein binding and cysteine-type peptidase activity	76.89	TINAGL1
AGT	Angiotensinogen	Renal tubular dysgenesis and hypertension, essential	Growth factor activity and serine-type endopeptidase inhibitor activity	65.57	Null
RHOC	Ras homolog family member C	Breast disease and breast cancer	GTP binding and obsolete signal transducer activity	64.59	RHOA
HP	Haptoglobin	Anhaptoglobinemia and *Plasmodium falciparum* malaria	Serine-type endopeptidase activity and hemoglobin binding	59.62	HPR
BCL2L1	BCL2-like 1	B-cell lymphomas and follicular lymphoma	Protein homodimerization activity and protein heterodimerization activity	49.62	BCL2
ABCC3	ATP binding cassette subfamily C member 3	Dubin-Johnson syndrome and extrahepatic cholestasis	Transporter activity and ATPase activity	46.90	ABCC1

expressed genes of each method and the number of unique co-differentially expressed genes excavated by each method. Unique co-differentially expressed genes selected by each method have been indicated in italic. In the table, we can clearly see that there are 0 unique co-differentially expressed genes excavated by the DNMF, LNMF, GNMF, and GDNMF methods; 2 unique co-differentially expressed genes excavated by the NMF and HNMF methods; and 7 unique co-differentially expressed genes mined by our HDNMF method. This indicates that our method works well.

In Table 7, we summarize the detailed information of the co-differentially expressed genes excavated only by the HDNMF method in the table by GeneCards, including the official name, related diseases, related GO annotations, RS, and paralog gene. All genes in the table have a high related score, which means that these genes can be regarded as important pathogenic genes to help disease research. Medical research in this area will also provide us with unexpected new discoveries. Therefore, the unique co-differentially expressed genes excavated by the HDNMF method can promote further research on these three diseases.

From Table 7, we can see the relevant information of the unique co-differentially expressed genes selected by the HDNMF method. ERBB3 has the highest relevance score. ERBB3 is a protein-coding gene. When it mutates, the tumor suppressor protein does not form properly. In pancreatic cancer, ERBB3 is a preferred dimerization partner for EGFR, and ERBB3 protein expression levels are directly related to the antiproliferative effects of erlotinib (EGFR-specific tyrosine kinase inhibitor) and transient knockdown of ERBB3 expression gain resistance to EGFR targeted therapy. This leads to the onset of cancer. ERBB3 has been shown to be associated with PAAD [33]. ERBB3 forms heterodimers with other kinase-active EGF (epidermal growth factor) receptor family members. Heterodimerization leads to activation of pathways leading to cell proliferation or differentiation. Amplification of this gene and/or overexpression of its protein has been reported in many cancers, such as PAAD, ESCA [34], and CHOL [35]. Therefore, mutations in one gene may be involved in the production of multiple cancers. This suggests that biologists can further study the link between different cancers.

3.4.2. INTA 2 Co-Differentially Expressed Gene Selection Results. Similar to Table 5, Table 8 was subjected to the same treatment to obtain NUM, TRS, and ARS of the INDA 2 co-differentially expressed genes. From Table 8, we can see that our method is better than other methods.

After the intersection of the co-differentially expressed genes selected by the five methods is removed, the co-differentially expressed genes excavated for each method are given in Table 9. As showed in Table 9, the number of unique public genes mined by the NMF, DNMF, LNMF, GNMF, GDNMF, and HNMF methods is 0, 0, 4, 0, 0, and 0. The HDNMF method has been tapped into 5. Next comparing the obtained co-differentially expressed genes, we can obtain unique co-differentially expressed genes obtained only by our method, and these genes are necessary for the study of diseases in the future. Therefore, our method is more suitable for mining co-differentially expressed genes. Unique co-differentially expressed genes selected by each method have been indicated in italic.

Some of the results from these and other studies are summarized below:

Table 8: NUM, TRS, and ARS of five methods (INTA 2).

Methods		HDNMF	NMF	DNMF	LNMF	GNMF	GDNMF	HNMF
PAAD	Num	326	326	*330*	321	321	325	319
	TRS	*2177.31*	2039.68	2164.29	2035.33	2037.44	1974.33	1945.6
	ARS	*6.6789*	6.2567	6.5585	6.3406	6.3472	6.0749	6.0991
ESCA	Num	263	265	*269*	261	263	269	253
	TRS	*2383.93*	2222.66	2274.52	2266.58	2170.57	2266.42	2298.05
	ARS	*9.0644*	8.3874	8.4555	8.6842	8.2531	8.4254	9.0832
HNSC	Num	245	*248*	247	240	247	248	242
	TRS	*5217.77*	4997	5065.33	5044.97	4954.63	5092.82	5117.22
	SRS	*21.2970*	20.1492	20.5074	21.0207	20.0592	20.5356	21.1455
COAD	Num	*341*	338	340	334	335	339	324
	TRS	*2134.17*	1995.88	2014.62	2027.45	1852.71	2010.03	1836.64
	ARS	*6.2586*	5.9050	5.9254	6.0702	5.5305	5.9293	5.6686
CHOL	Num	131	*136*	129	128	128	128	125
	TRS	*505.85*	493.98	455.1	466.04	424.14	464.75	402.63
	ARS	*3.8615*	3.6322	3.5280	3.6409	3.3136	3.6309	3.2210

Table 9: Co-differentially expressed genes of five methods (INTA 2).

Methods	Codifferentially expressed genes	Unique
NMF	CCND1 IFI27 FSCN1 LGALS3 RAC1 CLDN4 PRDX1 NOTCH3 COL1A1 HMGA1 EIF5A SERPINA3 POSTN HIF1A TP63 EGR1 SERPINB5 CDH3 KRT10 SPINK1 SLPI DUSP1	0
DNMF	KRT10 SPINK1 SERPINA3 CDH3 SERPINB5 TP63 SLPI HIF1A EGR1 ANXA5 DUSP1	0
LNMF	DMBT1 *LDHB TFF3 HMGB1 ATF4*	4
GNMF	SERPINA3 COL1A1 MUC1 KRT7 MMP7 MUC6 GRN SPINK1 AGR2	0
GDNMF	CCND1 IFI27 FSCN1 LGALS3 RAC1 CLDN4 PRDX1 NOTCH3 COL1A1 HMGA1 EIF5A POSTN SERPINA3 HIF1A TP63 EGR1 SERPINB5 CDH3 KRT10 SPINK1 SLPI DUSP1	0
HNMF	ERBB2 CCND1 MUC1 EPCAM LGALS3 AGR2 CLDN4 IFI27 FSCN1 RAC1 DMBT1 HMGA1 NOTCH3 CD151 PRDX1 EIF5A	0
HDNMF	AGR2 *AGT ALB ANPEP* ANXA5 CCND1 CD151 CLDN4 COL1A1 *CRP* DMBT1 EIF5A EPCAM ERBB2 *ERBB3* FSCN1 HMGA1 IFI27 KRT7 LGALS3 MMP7 MUC1 NOTCH3 PRDX1 RAC1 SERPINA3	5

(1) Unique co-differentially expressed genes mined by HDNMF on the INTA1 datasets are greater than the INTA 2 datasets. This huge difference is due to the complexity of the different datasets. For example, the composition of the INTA1 data has three diseases, and the INTA2 datasets constitute five diseases

(2) The genes AGT, ANPEP, CRP, and ERBB3 in Tables 8 and 10 have a higher correlation score with the experimental datasets, and these three genes are ignored by most mining methods. Therefore this reflects the accuracy of the HDNMF method

3.5. The Pathway Analysis. Joint biological processes of the selected genes selected in the experiment can be attributed to pathways that help us understand the advanced functions of biology and biological systems at the molecular level. We used the Kyoto Encyclopedia of Genes and Genomes (KEGG) online analysis tool to analyze co-differentially expressed genes identified by HDNMF. In this experiment, 500 genes identified were enclosed in KEGG, and we can obtain the corresponding disease pathway. The FDR of the HDNMF method and other methods are showed in Table 11 (INTA 1) and Table 12 (INTA 2). The smaller the FDR, the better the results. So, the HDNMF method achieves better results.

Taking ECM-receptor interaction as an example, the literature has been shown to be closely related to PAAD [36], ESCA [37], and CHOL [38]. Other pathways can also prove their rationality by consulting the literature.

4. Conclusions

We presented a novel matrix factorization method called Hypergraph Regularized Discriminative Nonnegative Matrix

Table 10: Unique co-differentially expressed genes by HDNMF (INTA 2).

Official gene symbol	Official name	Related diseases	Related GO annotations	RS	Paralog gene
ERBB3	Erb-B2 receptor tyrosine kinase 3	Retroperitoneal leiomyosarcoma and fatal congenital contracture syndrome 2	Protein homodimerization activity and transferase activity, transferring phosphorus-containing groups	94.06	ERBB4
CRP	C-reactive protein	Acute pancreatitis and appendicitis	Calcium ion binding and cholesterol binding	81.62	APCS
AGT	Angiotensinogen	Renal tubular dysgenesis and hypertension, essential	Growth factor activity and serine-type endopeptidase inhibitor activity	65.57	null
ALB	Albumin	Analbuminemia and hyperthyroxinemia, familial hypalbuminemia	Enzyme binding and chaperone binding	52.83	AFP
ANPEP	Alanyl aminopeptidase, membrane	Tetrasomy 21 and myelophthisic anemia	Virus receptor activity	40.34	LVRN

Table 11: The FDR of different methods (INTA 1).

Pathway	HDNMF	NMF	DNMF	LNMF	GNMF	GDNMF	HNMF
Ribosome	7.75E-91	3.16E-86	1.41E-102	9.94E-91	4.28E-93	*1.41E-102*	9.94E-91
Focal adhesion	*6.52E-38*	4.44E-35	1.8E-36	7.7E-38	1.8E-36	1.8E-36	7.7E-38
ECM-receptor interaction	*6.67E-36*	1.91E-32	1.91E-32	4.01E-34	1.91E-32	1.91E-32	4.01E-34
Pathogenic *Escherichia coli* infection	*8.01E-17*	8.33E-17	1.92E-15	1.7E-15	1.92E-15	1.92E-15	4.91E-17
Leukocyte transendothelial migration	8.01E-17	8.33E-17	1.04E-16	1.08E-18	1.04E-16	1.04E-16	*5.28E-18*

Table 12: The FDR of different methods (INTA 2).

Pathway	HDNMF	NMF	DNMF	LNMF	GNMF	GDNMF	HNMF
Ribosome	*3.15E-100*	1.74E-95	5.55E-84	1.98E-95	3.16E-86	4.28E-93	8.69E-98
Focal adhesion	*4.1E-35*	2.26E-32	2.43E-32	2.43E-32	2.43E-32	5.4E-31	2.43E-32
ECM-receptor interaction	*1.8E-32*	8.64E-31	9.13E-31	4.13E-29	9.13E-31	4.13E-29	4.13E-29
Pathogenic *Escherichia coli* infection	8.07E-17	*2.14E-18*	2.21E-18	2.21E-18	8.33E-17	2.21E-18	2.21E-18
Antigen processing and presentation	3.48E-15	2.9E-15	3E-15	*1.51E-16*	3E-15	1.51E-16	3E-15

Factorization (HDNMF). The method introduced hypergraph and discriminative label information into the standard NMF method. On the one hand, the hypergraph can find high-order geometric relations that are neglected by simple graphs; on the other hand, discriminative label information can make the method have supervisory functions. Experiments have shown that HDNMF can achieve better results than standard NMF and its improved methods. Since the hypergraph was in the process of construction, it got a lot of time to build the hyperedge; the program running time is longer than other methods. How to accelerate the proposed algorithm is an interesting work in the forthcoming work.

Data Availability

The datasets that support the findings of this study are available in https://cancergenome.nih.gov/.

Conflicts of Interest

There are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61872220 and 61572284.

References

[1] M. L. Metzker, "Sequencing technologies—the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[2] J. X. He and Y. F. Jiang, "The progress and prospect of application of genetic testing technology-based gene detection technology in the diagnosis and treatment of hereditary cancer," *Chinese Journal of Preventive Medicine*, vol. 51, no. 8, pp. 772–776, 2017.

[3] F. S. Tsai and K. I. Chan, "Dimensionality reduction techniques for data exploration," in *Proceedings of the International Conference on Information*, 2007.

[4] 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.

[5] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, pp. 556–562, 2001.

[6] C.-Y. Tsai and K.-T. Song, "Dynamic visual tracking control of a mobile robot with image noise and occlusion robustness," *Image and Vision Computing*, vol. 27, no. 8, pp. 1007–1022, 2009.

[7] M. M. Kalayeh, H. Idrees, and M. Shah, "NMF-KNN: image annotation using weighted multi-view non-negative matrix factorization," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 184–191, IEEE, Columbus, Ohio, USA, June 2014.

[8] J. Liu, D. Wang, Y. Gao, C. Zheng, Y. Xu, and J. Yu, "Regularized non-negative matrix factorization for identifying differentially expressed genes and clustering samples: a survey," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 3, pp. 974–987, 2018.

[9] V. L. Skrobot, E. V. R. Castro, R. C. C. Pereira, V. M. D. Pasa, and I. C. P. Fortes, "Use of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) in Gas Chromatographic (GC) Data in the Investigation of Gasoline Adulteration," *Energy & Fuels*, vol. 21, no. 6, pp. 5–19, 2007.

[10] X. Long, H. Lu, Y. Peng, and W. Li, "Graph regularized discriminative non-negative matrix factorization for face recognition," *Multimedia Tools and Applications*, vol. 723, pp. 2679–2699, 2014.

[11] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2691–2698, June 2010.

[12] K. Zeng, J. Yu, C. Li, J. You, and T. Jin, "Image clustering by hyper-graph regularized non-negative matrix factorization," *Neurocomputing*, vol. 138, pp. 209–217, 2014.

[13] Y. Peng, R. Tang, W. Kong, F. Qin, and F. Nie, "Parallel vector field regularized non-negative matrix factorization for image representation," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2018*, pp. 2216–2220, Canada, April 2018.

[14] Y. Peng, Y. Long, F. Qin, W. Kong, F. Nie, and A. Cichocki, "Flexible non-negative matrix factorization with adaptively learned graph regularization," in *Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3107–3111, Brighton, UK, May 2019.

[15] Y. Peng and B.-L. Lu, "Discriminative extreme learning machine with supervised sparsity preserving for image classification," *Neurocomputing*, vol. 261, pp. 242–252, 2017.

[16] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Huang, "Constrained nonnegative matrix factorization for image representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1299–1311, 2012.

[17] Y.-S. Yang, A.-B. Ming, Y.-Y. Zhang, and Y.-S. Zhu, "Discriminative non-negative matrix factorization (DNMF) and its application to the fault diagnosis of diesel engine," *Mechanical Systems and Signal Processing*, vol. 95, pp. 158–171, 2017.

[18] J. Yu, Y. Rui, and D. Tao, "Click prediction for web image reranking using multimodal sparse coding," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2019–2032, 2014.

[19] J. Yu, D. Tao, M. Wang, and Y. Rui, "Learning to rank using user clicks and visual features for image retrieval," *IEEE Transactions on Cybernetics*, vol. 45, no. 4, pp. 767–779, 2015.

[20] C. Hong, J. Yu, J. Wan, D. Tao, and M. Wang, "Multimodal deep autoencoder for human pose recovery," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5659–5670, 2015.

[21] M. R. Brito, E. L. Chávez, A. J. Quiroz, and J. E. Yukich, "Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection," *Statistics & Probability Letters*, vol. 35, no. 1, pp. 33–42, 1997.

[22] M. W. Jeter and W. C. Pye, "Some properties of Q-matrices," *Linear Algebra and its Applications*, vol. 57, pp. 169–180, 1984.

[23] P. Anttila, P. Paatero, U. Tapper, and O. Jarvinen, "Source identification of bulk wet deposition in Finland by positive matrix factorization," *Atmospheric Environment*, vol. 29, no. 14, pp. 1705–1718, 1995.

[24] P.-E. Danielsson, "Euclidean distance mapping," *Computer Graphics and Image Processing*, vol. 14, no. 3, pp. 227–248, 1980.

[25] C. Hong, J. Yu, J. Li, and X. Chen, "Multi-view hypergraph learning by patch alignment framework," *Neurocomputing*, vol. 118, no. 11, pp. 79–86, 2013.

[26] C. Wang, J. Yu, and D. Tao, "High-level attributes modeling for indoor scenes classification," *Neurocomputing*, vol. 121, pp. 337–343, 2013.

[27] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Proceedings of the 20th Annual Conference on Neural Information Processing Systems, NIPS 2006*, pp. 1601–1608, Canada, December 2006.

[28] J. Yu, D. Tao, and M. Wang, "Adaptive hypergraph learning and its application in image classification," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 21, no. 7, pp. 3262–3272, 2012.

[29] S. Wu, M. Hou, J. Liu, J. Wang, and S. Yuan, "Identifying characteristic genes and clustering via an l p-norm robust feature selection method for integrated data," in *Proceedings of the International Conference on Intelligent Computing*, pp. 419–431, Springer, 2018.

[30] O. Zoidi, A. Tefas, and I. Pitas, "Multiplicative update rules for concurrent nonnegative matrix factorization and maximum margin classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 3, pp. 422–434, 2013.

[31] D. Lu, Y. Sun, and S. Wan, "Brain tumor classification using non-negative and local non-negative matrix factorization," in *Proceedings of the 2013 IEEE International Conference on Signal Processing, Communications and Computing, ICSPCC 2013*, China, August 2013.

[32] S. Wu, M. Hou, J. Liu, J. Wang, and S. Yuan, "Identifying characteristic genes and clustering via an lp-norm robust feature selection method for integrated data," in *Intelligent Computing Theories and Application*, pp. 419–431, 2018.

[33] J. S. Liles, J. P. Arnoletti, C. D. Tzeng et al., "ErbB3 expression promotes tumorigenesis in pancreatic adenocarcinoma," *Cancer Biology & Therapy*, vol. 10, no. 6, pp. 555–563, 2010.

[34] H. Jiang, D.-P. Liu, D. Xie, and D.-Z. Wei, "Up-regulation of ErbB3-binding protein 1 inhibits the growth of esophageal carcinoma cells," *Sheng Li Xue Bao: [Acta Physiologica Sinica]*, vol. 68, no. 6, pp. 740–746, 2016.

[35] S.-I. Aishima, K.-I. Taguchi, K. Sugimachi, M. Shimada, K. Sugimachi, and M. Tsuneyoshi, "c-erbB-2 and c-Met expression relates to cholangiocarcinogenesis and progression of intrahepatic cholangiocarcinoma," *Histopathology*, vol. 40, no. 3, pp. 269–278, 2002.

[36] L. Zhao, T. Zhang, L. Zhuang, B. Yan, R. Wang, and B. Liu, "Retraction Note to: Uncovering the pathogenesis and identifying novel targets of pancreatic cancer using bioinformatics approach," *Molecular Biology Reports*, vol. 42, no. 10, pp. 1463-1463, 2015.

[37] D. Liu, "LYN, a key gene from bioinformatics analysis, contributes to development and progression of esophageal adenocarcinoma," *Medical Science Monitor Basic Research*, vol. 21, pp. 253–261, 2015.

[38] T. Xue, B. Zhang, S. Ye, and Z. Ren, "Differentially expressed gene profiles of intrahepatic cholangiocarcinoma, hepatocellular carcinoma, and combined hepatocellular-cholangiocarcinoma by integrated microarray analysis," *Tumor Biology*, vol. 36, no. 8, pp. 5891–5899, 2015.