Characteristic Gene Selection Based on Robust Graph Regularized Non-Negative Matrix Factorization

Dong Wang, Jin-Xing Liu, Ying-Lian Gao, Chun-Hou Zheng, and Yong Xu

Abstract—Many methods have been considered for gene selection and analysis of gene expression data. Nonetheless, there still exists the considerable space for improving the explicitness and reliability of gene selection. To this end, this paper proposes a novel method named robust graph regularized non-negative matrix factorization for characteristic gene selection using gene expression data, which mainly contains two aspects: Firstly, enforcing L_{21}-norm minimization on error function which is robust to outliers and noises in data points. Secondly, it considers that the samples lie in low-dimensional manifold which embeds in a high-dimensional ambient space, and reveals the data geometric structure embedded in the original data. To demonstrate the validity of the proposed method, we apply it to gene expression data sets involving various human normal and tumor tissue samples and the results demonstrate that the method is effective and feasible.

Index Terms—Gene expression data, gene selection, manifold embed, $L_{2,1}$-norm, nonnegative matrix factorization

1 Introduction

Tumor is one of the most harmful diseases, which has plagued the human for many years. It is said that tens of thousands of people die from tumor every year. Therefore it is important and challenge for scientists to find the virulence genes in numerous gene expression data. With the development of microarray technologies [1], [2], [3], [4], many methods have been used in gene expression data for gene selection [5], [6], [7]. For example, Journée et al. proposed a sparse principal component analysis (SPCA) method by using generalized power method [8]. Witten et al. proposed a penalized matrix decomposition (PMD) method which has been expressed useful in microarray analysis via imposing penalization on factor matrices [9]. Nonnegative matrix factorization with sparse constraints (NMFSC), which was first introduced by Hoyer in 2004 [10], has been widely used in gene selection. Devarajan demonstrate that NMF is an analytical and interpretive tool in computational biology [11]. Brunette et al. using matrix factorization discovery meta-genes and molecular pattern [12]. Utro and Giancarlo proposed the Consensus Clustering methodology for microarray data analysis [13].

Though the aforementioned methods are useful, there are two characteristics of gene expression data that present challenging problems for traditional machine-learning methods: Firstly, in many real data applications, we usually consider the samples in low-dimensional manifold which embeds in a high-dimensional ambient space, and thus, it is important and necessary to consider the data geometric structure embedded in the original data. However, these methods do not consider that. Secondly, the real gene expression data often contain many outliers and noises, however these methods cannot deal with the outliers and noises effectively.

Facing with the first problem and from the geometric perspective, it is obvious that manifold learning is an appropriate method to consider the data geometric structure embedded in the original data. Cai et al. proposed graph regularized nonnegative matrix factorization (GNMF) [14], [15] model to preserve geometrical information by constructing affinity graph but it cannot deal with the outliers and noises. The second problem can be solved by enforcing L_{21}-norm on object function [16].

Motivated by previous researches [17], [18], in this paper, we propose a novel method, robust graph regularized nonnegative matrix factorization (RGNMF) algorithm, via imposing the Manifold Regularization to discover the low dimension manifold embedded in a high dimensional ambient space and enforcing $L_{2,1}$-norm [19] instead of L_2-norm on error function to diminish the impact of noisy and outliers [20], [21], [22].

The mainly contribution of this paper is described as follows:

- *D. Wang is with the School of Information Science and Engineering, Qufu Normal University, Qufu, Shandong 276826, China. E-mail: dongwshark@126.com.*
- *J.-X. Liu is with the School of Information Science and Engineering, Qufu Normal University, Qufu, Shandong 276826, China, and with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, Guangdong 518055, China. E-mail: sdcavell@126.com.*
- *Y.-L. Gao is with the Library of Qufu Normal University, Qufu Normal University, Qufu, Shandong 276826, China. E-mail: yinliangao@126.com.*
- *C.-H. Zheng is with the College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui 230039, China. E-mail: zhengch99@126.com.*
- *Y. Xu is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, and with the Key Laboratory of Network Oriented Intelligent Computation, Shenzhen, Guangdong 518055, China. E-mail: yongxu@ymail.com.*

Manuscript received 19 July 2015; revised 0 . 0000; accepted 23 Nov. 2015. Date of publication 8 Dec. 2015; date of current version 5 Dec. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2015.2505294

Firstly, $L_{2,1}$-norm based error function is used to reduce the outliers and noises in real gene expression data.

Secondly, the addition of manifold regularization aims to find the low dimensional manifold in the high dimensional feature space and can find the inherent law of data from the observations [23].

The rest of the paper is organized as follows. In Section 2, we propose the RGNMF method and demonstrate the efficient algorithm of our method. In Section 3, we compare our RGNMF method with other four methods (GNMF, NMFSC, PMD and SPCA). Finally, the conclusions are given in Section 4.

2 Methodology

We begin with the study of L_{21}-norm, followed by a robust feature selection which unifies L_{21}-norm and manifold based on NMF method. Then we give an efficient algorithm to solve the minimization problem.

2.1 Mathematical Definition of $L_{2,1}$

This subsection briefly introduces the $L_{2,1}$-norm which was proposed in [21]. It is defined as

$$\|\mathbf{M}\|_{2,1} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \sqrt{\sum_{j=1}^{\mathrm{s}} \mathbf{m}_{ij}^2} = \sum_{i=1}^{n} \left\|\mathbf{m}^i\right\|_2, \tag{1}$$

here $\mathbf{m}^i$ is the ith row of $\mathbf{M}$. The explanation of $L_{2,1}$-norm is that we firstly compute L_2-norm of rows $\mathbf{m}^i$ and then compute L_1-norm of vector . The amplitude of components of vector $\mathrm{b}(\mathbf{M})$ dedicates how important each dimension is. $L_{2,1}$-norm favors a small number of non-zero rows in $\mathbf{M}$, accordingly ensuring dimension reduction to be achieved [20], [24].

2.2 Robust Graph regularized Non-Negative Matrix Factorization

We first introduce the standard NMF algorithm in our method. Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_j) \in R^{\mathrm{i}\times j}$, $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ...\mathbf{y}_j) \in R^{d\times j}$, the error function of standard NMF [25] is

$$\|\mathbf{X} - \mathbf{AY}\|_F^2 = \sum_{n=1}^{j} \left\|\mathbf{x}_n - \mathbf{A}\mathbf{y}_n\right\|^2, \quad st. \mathbf{Y} > \mathbf{0}, \mathbf{A} > \mathbf{0}. \tag{2}$$

Here the error for each data point enters the objective function as squared residue error in a manner of $\|\mathbf{x}_n - \mathbf{A}\mathbf{y}_n\|^2$. As a result, a few large outliers and errors can easily dominate the objection function because of the squared errors. So, it is necessary to propose the $L_{2,1}$-norm formulation to reduce the outliers [21], [26].

The development of graph theory and manifold learning theory [27] have demonstrate that the local geometric structure embedded in high dimension can be effectively modeled through a nearest neighbor graph. For each data point x_i we find its k nearest neighbors and put edges between and its neighbors. There are many choices to define the weight matrix $\mathbf{W}$ on the graph. The w_{jl} is used to measure the closeness of two points x_j andx_l.

The method to defined the weight matrix is as follows:

0-1 weight: $w_{jl} = 1$, if and only if node j and l are connected by edge. This is the simplest weighting method and is very easy to compute.

The $w_{jl} = 1$ is only for measuring the closeness of two data point in this method, so we apply the Euclidean distance $O(s_j, s_l) = \|s_j - s_l\|^2$ to measure the distance between the low dimension of two data points.

Then the smoothness of the low dimension representation can be measured by:

$$\begin{aligned} R &= \frac{1}{2}\sum_{j,l}^{N} \left\|s_j - s_l\right\|^2 w_{jl} \\ &= \sum_{j=1}^{N} s_j^T s_j d_{\mathrm{jj}} - \sum_{j,l=1}^{N} s_j^T s_l w_{jl} \\ &= \mathrm{Tr}(\mathbf{YDY}^T) - \mathrm{Tr}(\mathbf{YWY}^T) \\ &= \mathrm{Tr}(\mathbf{YLY}^T). \end{aligned} \tag{3}$$

The error function of the RGNMF formulation is

$$\min_{\mathbf{A},\mathbf{Y}} \|\mathbf{X} - \mathbf{AY}\|_{2,1} + \lambda \mathrm{Tr}(\mathbf{YLY}^T), \quad st. \mathbf{Y} > \mathbf{0}, \mathbf{A} > \mathbf{0}, \tag{4}$$

where $\mathrm{Tr}(\cdot)$ denotes a trace of matrix, the regularization parameter $\lambda \geq 0$ controls the smoothness of the new representation [28]. $\mathbf{W}$ is a weight matrix of the nearest neighbor graph and $\mathbf{D}$ is a diagonal matrix whose entries are column (or row, since $\mathbf{W}$ is symmetric) sums of $\mathbf{W}$, $d_{jj} = \sum_c w_{jc}$. $\mathbf{L} = \mathbf{D} - \mathbf{W}$, which is called graph Laplacian [29].

2.3 An Efficient Algorithm of RGNMF

The error function in Eq. (4) can be rewritten as

$$\begin{aligned} f &= \mathrm{Tr}((\mathbf{X} - \mathbf{AY})\mathbf{G}(\mathbf{X} - \mathbf{AY})^T) + \alpha \mathrm{Tr}(\mathbf{Y}^T\mathbf{LY}) \\ &= \mathrm{Tr}(\mathbf{XGX}^T) - 2\mathrm{Tr}(\mathbf{XGY}^T\mathbf{A}^T) + \mathrm{Tr}(\mathbf{AYGY}^T\mathbf{A}^T) \\ &\quad + \alpha \mathrm{Tr}(\mathbf{YLY}^T), \end{aligned} \tag{5}$$

where $\mathbf{G}$ is a diagonal matrix with diagonal elements given by

$$\mathbf{G}_{jj} = 1/\sqrt{\sum_{m=1}^{i} (\mathbf{X} - \mathbf{AY})_{mj}^2 + \varepsilon} = 1/\left\|x_j - \mathbf{F}g^j + \varepsilon\right\|, \tag{6}$$

where ε is a positive number and infinitely close to but not equal to zero.

To solve the constrained optimization problem in Eq. (4), we first introduce Lagrange multipliers ψ_{ik} and ϕ_{kj} by constraint $a_{ik} \geq 0$ and $y_{kj} \geq 0$, respectively [30]. Let $\Psi = [\psi_{ik}]$ and $\Phi = [\phi_{kj}]$, the Lagrange function L is defined as

$$\begin{aligned} L = \mathrm{Tr}(\mathbf{XGX}^T) &- 2\mathrm{Tr}(\mathbf{XGY}^T\mathbf{A}^T) + \mathrm{Tr}(\mathbf{AYGY}^T\mathbf{A}^T) \\ &+ \lambda \mathrm{Tr}(\mathbf{YLY}^T) + Tr(\Psi\mathbf{A}^T) + Tr(\Phi\mathbf{Y}^T). \end{aligned} \tag{7}$$

Using the KKT conditions [31] $\psi_{ik}a_{ik} = 0$ and $\phi_{kj}y_{\mathrm{kj}} = 0$, left multiplying the two sides of the derivatives of L with respect to $\mathbf{A}$ and $\mathbf{Y}$ by a_{ik} and y_{kj}, we obtain the updating rules which are listed as follows:

$$a_{ik} \leftarrow a_{ik} \frac{(\mathbf{XGY}^T)_{ik}}{(\mathbf{AYGY}^T)_{ik}}, \tag{8}$$

$$y_{kj} \leftarrow y_{kj} \frac{(\mathbf{A}^T\mathbf{XG} + \lambda\mathbf{YW})_{kj}}{(\mathbf{A}^T\mathbf{AYG} + \lambda\mathbf{YD})_{kj}}. \tag{9}$$

The detail of our method is summarized as listed in Algorithm 1. The iteration procedure is repeated until the algorithm converges.

The advantage of multiplicative updating rules is guarantee of non-negative of $\mathbf{A}$ and $\mathbf{Y}$.

Algorithm 1. RGNMF

Input: $\mathbf{X} \in R^{i\times j}$ and parameter λ.
Output: $\mathbf{Y} \in R^{k\times j}$,$\mathbf{A} \in \mathbf{R}^{i\times k}$,$\mathbf{W} \in \mathbf{R}^{j\times j}$
1: Initialize $\mathbf{A}_0 \in \mathbf{R}^{i\times k}$ and$\mathbf{Y}_0 \in \mathbf{R}^{k\times j}$ as non-negative matrices, $\mathbf{G}_0 \in R^{j\times j}$ as an identity matrix.
Set r = 0.
2: **repeat**
Update $\mathbf{A}_{r+1}$ as
$\mathbf{A}_{r+1} \leftarrow \mathbf{A}_r \frac{\mathbf{X}\mathbf{G}_r\mathbf{Y}_r^T}{\mathbf{A}_r\mathbf{Y}_r\mathbf{G}_r\mathbf{Y}_r^T}$
Update $\mathbf{Y}_{r+1}$ as
$\mathbf{Y}_{r+1} \leftarrow \mathbf{Y}_r \frac{(\mathbf{A}^T{}_{r+1}\mathbf{X}\mathbf{G}_{r+1}+\lambda\mathbf{Y}_r\mathbf{W})_{jk}}{(\mathbf{A}^T{}_{r+1}\mathbf{A}\mathbf{Y}_r\mathbf{G}_{r+1}+\lambda\mathbf{Y}_r\mathbf{D})_{jk}}$
Compute diagonal matrix $\mathbf{G}_{r+1}$ according to Eq. (7).
r = r+1
Until convergence

Theorem 1. *The object function in Eq. (4) is non-increasing under the updating rules in Eq. (8) and (9).*

Please see the Appendix for a detailed proof of the above theorem. Our proof essentially follows the idea in the proof of Lee and Seung's paper [32].

2.4 Extracting Characteristic Genes by RGNMF

In this paper, based on $L_{2,1}$-norm and manifold regularization algorithm, a new method is introduced to gain the differentially expressed genes. The gene expression data are gathered in the matrix $\mathbf{X}$ with size $i \times j$. Each row of $\mathbf{X}$ represents the transcriptional response of the ith genes in all sample. So, the nonnegative matrix factorization of $\mathbf{X}$ can be written as:

$$\mathbf{X} \approx \mathbf{A}\mathbf{Y}, \tag{10}$$

where $\mathbf{Y}$ is a $k \times j$ matrix which is called the coefficient matrix, $\mathbf{A}$ is an $i \times k$ matrix which is called the basis matrix, among them $k \leq \min(i, j)$. Then, we can extract differentially expressed genes from the basis matrix $\mathbf{A}$, the progress is listed in the following:

The matrix $\mathbf{A}$ can be described as follows:

$$\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K]^T. \tag{11}$$

We can get the evaluating vector $\mathbf{A}$ which is sorted in descending order:

$$\mathbf{A} = \left[\sum_{m=1}^{k} |\mathbf{A}_{1m}|, \ldots, \sum_{m=1}^{k} \left|\mathbf{A}_{jm}\right|\right]^T. \tag{12}$$

Without loss of generality, the larger the element in $\mathbf{A}$ is, the more differential the gene is. So, the genes associated with the first num ($\mathrm{num} \leq j$) largest entries in $\mathbf{A}$ are selected as differentially expressed ones.

As a conclusion, we describe the RGNMF method to extract differentially expressed genes as follows:

1) Gain the data matrix $\mathbf{X}$ according to gene expression data.
2) Obtain the diagonal matrix $\mathbf{W}$ and basis matrix $\mathbf{A}$ by using RGNMF method.
3) Extract the differentially expressed genes via matrix $\mathbf{A}$.
4) Check the extracted genes through gene ontology (GO) tool.

3 Results and Discussion

In order to evaluate the performance of RGNMF method, several experiments are carried out to compare our method with the following four ones: (a) GNMF method (proposed by Cai et al. [14]); (b) NMFSC method (proposed by Hoyer [33]); (c) SPCA method (proposed by Journée et al. [34]); (d) PMD method (proposed by Witten et al. [35]). We perform these methods on four publicly available datasets, i.e., leukemia dataset [36], medulloblastoma dataset [37], diffuse large B cell lymphoma (DLBCL) dataset [38], and lung carcinomas dataset [39].

All the parameters we get from the comparison algorithm are following the results described in their own papers. For example, GNMF algorithm, which used graph regularization to preserve the local structures of data, needs to construct a K-nearest neighbor graph. We set the distance metric as Euclidean distance, the value of K as 5, and the mode as the heat kernel in LPP [40].

3.1 Gene Ontology Analysis

In this paper, the tool to evaluate differentially expressed genes is GO Term which can provides profound information for the biological interpretation of high-throughput experiments. The gene ontology enrichment of functional annotation of the extracted genes by five methods is detected by ToppFun which is publicly available at: http://toppgene.cchmc.org/enrichment.jsp.

For a fair comparison, 100 genes are selected from gene expression data by GNMF, NMFSC, SPCA, PMD and RGNMF methods.

3.2 Leukemia Dataset

The leukemia dataset has become a reference in tumor selection. In this dataset, the distinction between acute myelogenous leukemia (AML) and acute lymphoblastic leukemia (ALL), as well as the division of ALL into T and B cell subtypes, is known. The dataset contains p = 5000 genes in 38 samples, and consists of 19 cases of B cell ALL (ALL_B), 8 cases of T cell ALL (ALL_T), and 11 cases of AML [36].

Table 1 lists the 10 closely related to the leukemia terms of P-values corresponding to different methods. In this table, the lowest P-values among the five methods is marked in bold type. It can be found that our method performs better than GNMF in seven terms. In another three terms (GO:0042277, GO:0033218, GO:0003823), our method has the same P-value with the GNMF. Only in one term (GO:0001772), NMFSC performs better than our method and it is obvious that our method is superior to PMD and SPCA.

Inspired by the work, there are 50 genes that are most highly correlated with the ALL-AML class distinction. Among the 50 genes, 35 genes (the total number of selected genes are 1,000) are selected by RGNMF, part of them are listed in Table 2. In the top 30 genes in the selected 100 genes, there are 28 genes are also selected in [41].

3.3 Medulloblastoma Dataset

We also apply these methods to the Medulloblastoma dataset [37], which is about childhood brain tumors. The pathogenesis of these tumors is not well understood, but it is

TABLE 1
The Leukemia Terms of P-Values Corresponding to Different Methods

ID	Name	RGNMF	GNMF	NMFSC	SPCA	PMD
17092989-SuppTable1	Human Lymphoma Fogel07 33genes	**4.95E-38**	6.63E-35	2.30E-27	1.14E-17	9.91E-23
18689800-TableS7	Human Embryonic Stem Cell Thomas08 1088genes	**5.05E-20**	1.12E-18	3.30E-17	1.78E-10	2.36E-12
12086872-Table 11c	Human Leukemia Yeoh02 40genes T-ALL	**3.17E-18**	2.55E-17	7.36E-16	1.55E-15	2.23E-14
GO:0042277	peptide binding	**2.55E-10**	**2.55E-10**	1.53E-04	1.53E-08	6.63E-07
GO:0033218	amide binding	**3.92E-10**	**3.92E-10**	1.90E-04	none	9.01E-07
GO:0003823	antigen binding	**6.71E-08**	**6.71E-08**	2.00E-06	3.05E-06	2.32E-05
GO:0032403	protein complex binding	**6.00E-07**	8.54E-05	**1.90E-07**	2.88E-06	1.04E-05
GO:0005615	extracellular space	**4.24E-07**	2.11E-06	7.04E-05	1.73E-03	6.02E-04
GO:0016023	cytoplasmic membrane-bounded vesicle	**8.35E-07**	4.35E-06	1.96E-03	none	none
GO:0001772	immunological synapse	3.43E-06	1.53E-04	**7.24E-08**	1.02E-07	3.04E-06

'none' denotes that the algorithm cannot give the GO terms.

generally accepted that there are two known histological subclasses: classic and desmoplastic, whose differences can be clearly seen under the microscope. In our experiment, the dataset contains $p = 5{,}893$ genes in 34 samples. The samples can be divided into 25 classic and nine desmoplasitic medulloblastomas. We analyze 34 medulloblastoma samples whose histology was scored by using World Health Organization (WHO) criteria.

Table 3 lists the 10 closely related to the medulloblastoma terms of P-values corresponding to different methods. In this table, the lowest P-values among the five methods are marked in bold type. It can be found that RGNMF outperforms the others in all the terms.

A number of genes not previously associated with clinical outcome were identified. There are fifty genes most highly associated with favorable outcome and treatment failure according to the signal-to-noise metric. These correlated with favorable outcome include many genes characteristic of cerebellar differentiation (vesicle coat protein b-NAP, NSCL1, TRKC, sodium channels), and genes encoding extracellular matrix proteins (PLOD lysyl hydroxylase, collagen type V elastin). As expected, TRKC expression is correlated with a favorable outcome, consistent with previous reports of this association [42]. Among the 100 genes, 51 genes (the total number of selected genes are 1,000) are selected by RGNMF, part of them are listed in Table 4. In the top 30 genes in the selected genes, there are eight genes are also selected in [43].

3.4 Diffuse Large B Cell Lymphoma Dataset

Diffuse large B-cell lymphoma, the most common lymphoid malignancy in adults, is curable in less than 50 percent of patients. Shipp et al. applied a supervised learning method on an expression profiling dataset of 7,139 genes on 58 tumor specimens, and identified 13 genes that are highly predictive to the outcomes [38]. The research presented in this paper uses Shipp's dataset [www.genome.wi.mit.edu/MPR/lymphoma].

TABLE 2
Leukemia Genes Extracted by RGNMF

Gene ID	Accession NO	Gene name	Gene function
100	M13792	ADA	It encodes an enzyme that catalyzes the hydrolysis of adenosine to inosine.
4792	M69043	MAD-3	It encodes a member of the NF-kappa-B inhibitor family, which contain multiple ankrin repeat domains.
7155	Z15115	TOP2B	It encodes a DNA topoisomerase, an enzyme that controls and alters the topologic states of DNA during transcription.
5687	X59417	PSMA6	It encodes a member of the peptidase T1A family, that is a 20S core alpha subunit.
896	M92287	CCND3	This protein has been shown to interact with and be involved in the phosphorylation of tumor suppressor protein Rb.
4602	U22376	MYB	It encodes a transcription factor that is a member of the MYB family of transcription factor genes. This protein plays an essential role in the regulation of hematopoiesis and may play a role in tumorigenesis.
973	U05259	CD79A	It encodes the Ig-alpha protein of the B-cell antigen component. The B lymphocyte antigen receptor is a multimeric complex that includes the antigen-specific component, surface immunoglobulin (Ig). Surface Ig non-covalently associates with two other proteins, Ig-alpha and Ig-beta, which are necessary for expression and function of the B-cell antigen receptor.
5552	X17042	SRGN	It encodes a protein best known as a hematopoietic cell granule proteoglycan. This encoded protein was found to be associated with the macromolecular complex of granzymes and perforin, which may serve as a mediator of granule-mediated apoptosis.
3576	Y00787	CXCL8	It is believed to play a role in the pathogenesis of bronchiolitis, a common respiratory tract disease caused by viral infection.
4069	M19045	LYZ	It encodes human lysozyme, whose natural substrate is the bacterial cell wall peptidoglycan.
1509	M63138	CTSD	It encodes a lysosomal aspartyl protease composed of a dimer of disulfide-linked heavy and light chains.

TABLE 3
The Medulloblastoma Terms of P-Values Corresponding to Different Methods

ID	Name	RGNMF	GNMF	NMFSC	SPCA	PMD
GO:0006415	translational termination	**8.13E-84**	1.23E-77	7.19E-78	9.59E-68	6.91E-56
GO:0022626	cytosolic ribosome	**6.08E-84**	2.20E-80	8.62E-81	7.30E-68	3.03E-56
GO:0006414	translational elongation	**9.08E-81**	1.47E-75	1.86E-77	2.61E-66	2.04E-57
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	**9.08E-81**	6.08E-75	5.26E-75	2.18E-65	6.38E-54
GO:0006613	cotranslational protein targeting to membrane	**2.64E-80**	1.56E-74	1.44E-74	5.00E-65	1.28E-53
GO:0044391	ribosomal subunit	**5.49E-78**	1.05E-70	1.54E-70	7.75E-62	1.97E-51
GO:0003735	structural constituent of ribosome	**1.00E-71**	4.07E-67	3.35E-67	8.85E-59	2.40E-48
GO:0005840	ribosome	**2.52E-69**	3.00E-65	1.46E-64	4.64E-55	1.17E-45
M4481	Genes up-regulated in comparison of naive B cells versus un-stimulated neutrophils.	**8.82E-35**	3.80E-33	1.88E-32	1.93E-29	9.76E-25
M3327	Genes down-regulated in comparison of poly some bound (translated) mRNA versus total mRNA in dendritic cells.	**5.66E-28**	3.37E-26	7.47E-26	9.20E-28	4.04E-23

Table 5 lists the eight closely related to DLBCL terms of P-values corresponding to different methods. In this table, the lowest P-values among the five methods are marked in bold type. It can be found that our method performs better than GNMF, SPCA and PMD in all terms. In M19251 and 15790779-Table 2 terms, our method has the same P-value with the NMFSC.

3.5 Lung Carcinomas Dataset

Carcinoma of the lung claims more than 150,000 lives every year in the United States. More fundamental knowledge of the molecular basis of lung carcinomas could aid in the prediction of patient outcome, the informed selection of currently available therapies, and the identification of novel molecular targets for chemotherapy [44].

TABLE 4
Medulloblastoma Genes Extracted by RGNMF

Gene ID	Accession NO	Gene name	Gene function
4807	M96739	NSCL1	The helix-loop-helix (HLH) proteins are a family of putative transcription factors.
1400	D78012	CRMP1	It encodes a member of a family of cytosolic phosphoproteins expressed exclusively in the nervous system.
2778	M21142	GNAS	This locus has a highly complex imprinted expression pattern.
166	U04241	AES	The protein encoded by this gene is similar in sequence to the amino terminus of Drosophila enhancer of split groucho.
6222	X69150	RPS18	It is an ortholog of mouse Ke3.
6130	M36072	RPL7A	It rearranges with the trk proto-oncogene to form the chimeric oncogene trk-2h.
6204	U14972	RPS10	The protein belongs to the S10E family of ribosomal proteins.
4736	U12404	HSPB1	It encodes a ribosomal protein that is a component of the 60S subunit. The protein belongs to the L1P family of ribosomal proteins.
6633	U15008	SNRPD2	The protein encoded by this gene belongs to the small nuclear ribonucleoprotein core protein family.
6159	Z49148	RPL29	The protein belongs to the L29E family of ribosomal proteins.
6602	X67247	RPS8	It is co-transcribed with the small nucleolar RNA genes.
6157	U14968	RP L27a	It encodes a ribosomal protein that is a component of the 60S subunit.
312962	HG613-HT613	RPS12	It encodes a ribosomal protein that is a component of the 40S subunit.
6223	M81757	RPS19	It encodes a ribosomal protein that is a component of the 40S subunit. The protein belongs to the S19E family of ribosomal proteins.
2023	M14328	ENO1	It encodes alpha-enolase, one of three enolase isoenzymes found in mammals. Alternative splicing of this gene results in a shorter isoform that has been shown to bind to the c-myc promoter and function as a tumor suppressor.
471	D82348	ATIC	It encodes a ribosomal protein that is a component of the 60S subunit. The protein belongs to the L15P family of ribosomal proteins.
6925	HG2479-HT2575	TCF4	It is broadly expressed, and may play an important role in nervous system development. Defects in this gene are a cause of Pitt-Hopkins syndrome.
6165	X52966	RPL35A	It encodes a ribosomal protein that is a component of the 60S subunit. The protein belongs to the L35AE family of ribosomal proteins.
6135	X79234	RPL11	It encodes a ribosomal protein that is a component of the 60S subunit. The protein belongs to the L5P family of ribosomal proteins.
3921	M14199	LAMR1	It encodes a high-affinity, non-integrin family, laminin receptor 1.

TABLE 5
The DLBCL Terms of P-values Corresponding to Different Methods

ID	Name	RGNMF	GNMF	NMFSC	SPCA	PMD
M11197	Housekeeping genes identified as expressed across 19 normal tissues.	**2.70E-100**	1.09E-96	1.72E-98	1.96E-87	6.86E-95
M19251	Genes up-regulated in Caco-2 cells after coculture with the probiotic bacteria L. casei for 6h.	**3.54E-46**	5.74E-44	**2.27E-46**	6.10E-24	1.37E-41
16872506-SuppTable1	Human Leukemia Yukinawa06 2000genes	**1.48E-44**	2.04E-41	3.95E-43	1.22E-30	1.22E-43
18931459-Table 2	Human Leukemia Lee08 66genes	**1.06E-41**	5.44E-37	2.47E-39	5.17E-16	3.78E-37
15790779-Table 2	Human BoneMarrow Jeong04 50genes	**6.07E-38**	1.05E-35	**4.28E-38**	1.37E-14	2.08E-35
M2535	The '3/3 signature': genes consistently down-regulated in all three pools of normal mammary stem cells.	**5.67E-31**	2.32E-29	2.75E-27	1.07E-22	1.58E-29
16651414-Supp2	Human Breast Bertucci06 2537genes	**1.36E-31**	1.76E-30	1.25E-30	7.57E-19	1.87E-30
M15774	Genes whose promoters are bound by MYC, according to MYC Target Gene Database.	**7.27E-16**	4.25E-15	4.06E-15	7.09E-12	2.54E-15

A total of 203 snap-frozen lung tumors (n = 186) and normal lung (n = 17) specimens were used to create two datasets. Of these, 125 adenocarcinoma samples were associated with clinical data and with histological slides from adjacent sections. The 203 specimens include histologically defined lung adenocarcinomas (n = 127), squamous cell lung carcinomas (n = 21), pulmonary carcinoids (n = 20), SCLC (n = 6) cases, and normal lung (n = 17) specimens. Other adenocarcinomas (n = 12) were suspected to be extrapulmonary metastases based on clinical history (see SampleData.xls, which is published as supporting information on the PNAS web site, www.pnas.org, and at www.genome.wi.mit.edu MPR lung).

Table 6 lists the nine closely related to lung carcinomas terms of P-values corresponding to different methods. From the multi-class dataset, it can be found that our method performs better than others in all terms.

4 Conclusions

In this paper, we propose an effective method with enforcing $L_{2,1}$-norm and Graph Regularized on error function. Accordingly, by previous works, the $L_{2,1}$-norm can diminish the outliers and noisy data, manifold regularization can find the low dimensional manifold in the high dimensional feature space, and calculate the corresponding embedding mapping, and find the inherent law of data from the observations. We also use the non-negative factorization method to avoid the high dimension, non-negative problems. To the summary, our method can handle high-dimension, non-negative, outliers and manifold embed simultaneously.

Furthermore, the genes selected by the methods on four tumor datasets of gene expression are analyzed by using GO terms enrichment. The results indicate that the proposed RGNMF method has superiority over GNMF, NMFSC, SPCA and PMD on extracting the differentially expressed genes.

COMPETING INTERESTS

The authors declare that they have no competing interests.

Appendix

Proofs of Theorem

To prove Theorem 1, we need to show that the object function in Eq. (4) is non-increasing under the updating rules. For the objective function, we need to fix **A** if we update **Y**. Similarly, we need to fix **Y** if we update **A**. Hence, we only need to prove that Eq. (4) is non-increasing under the update rules. Our proof make use of an auxiliary function similar to that used in the Expectation-Maximization algorithm [45]. We first give the definition of the auxiliary function [18].

TABLE 6
The Lung Carcinomas Terms of P-values Corresponding to Different Methods

ID	Name	RGNMF	GNMF	NMFSC	PMD	SPCA
GO:0006415	translational termination	**9.58E-19**	6.38E-17	7.55E-17	1.38E-15	2.34E-03
GO:0022626	cytosolic ribosome	**8.43E-19**	5.19E-17	5.21E-17	1.03E-15	1.23E-03
GO:0006613	cotranslational protein targeting to membrane	**7.20E-18**	2.31E-17	1.94E-17	1.91E-16	5.39E-03
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	**5.49E-18**	3.58E-16	3.05E-16	6.06E-15	6.38E-02
M11197	Housekeeping genes identified as expressed across 19 normal tissues.	**1.47E-18**	2.58E-18	2.58E-18	1.37E-17	1.16E-05
GO:0006414	translational elongation	**2.29E-18**	1.38E-16	1.54E-16	1.89E-15	none
GO:0005840	ribosome	**1.27E-16**	1.65E-16	2.03E-16	1.34E-15	2.58E-02
GO:0003735	structural constituent of ribosome	**1.39E-15**	1.77E-15	1.94E-15	2.32E-14	1.23E-02
GO:0005198	poly(A) RNA binding	**4.05E-10**	5.04E-10	4.86E-10	1.13E-07	none

'none' denotes that the algorithm cannot give the GO terms.

Definition. *$G(y, y')$ is an auxiliary function of $F(y)$ if the conditions*

$$G(y, y') \geq F(y), G(y, y) = F(y). \tag{13}$$

This auxiliary function is very useful because of the following lemma.

Lemma 1. *If G is an auxiliary function of F, then F is non-increasing under the update*

$$y^{(t+1)} = \arg\min_{h} G(y, y^{(t)}). \tag{14}$$

Proof.

$$F(y^{(t+1)}) \leq G(y^{(t+1)}, y^{(t)}) \leq G(y^{(t)}, y^{(t)}) = F(y^{(t)}). \quad \square$$

Now we will show that the update step for $\mathbf{Y}$ in Eq. (9) is exactly the update in Eq. (14) with a proper auxiliary function.

Considering any element y_{ba} in $\mathbf{Y}$, we use F_{ba} to denote the part of Eq. (4) which is only relevant to y_{ba}. It is easy to obtain the following derivatives:

$$\begin{aligned} F'_{ba} = \left(\frac{\partial O_1}{\partial \mathbf{Y}}\right)_{ba} &= (-2\mathbf{A}^T\mathbf{X}\mathbf{G} + 2\mathbf{A}^T\mathbf{A}\mathbf{Y}\mathbf{G} + 2\lambda\mathbf{Y}\mathbf{L})_{ba} \\ F''_{ba} &= 2(\mathbf{A}^T\mathbf{A}\mathbf{G})_{bb} + 2\lambda\mathbf{L}_{aa}. \end{aligned} \tag{15}$$

Since our update is essentially wised, it is essential to show that each F_{ba} is non-increasing under the update rules. Consequently, we introduce the following lemma.

Lemma 2. *Function*

$$\begin{aligned} G(y, y_{ba}^{(t)}) = F_{ba}(y_{ba}^{(t)}) &+ F'_{ba}(y_{ba}^{(t)})(y - y_{ba}^{(t)}) \\ &+ \frac{(\mathbf{A}^T\mathbf{A}\mathbf{Y}\mathbf{G})_{ba} + \lambda(\mathbf{Y}\mathbf{D})_{ba}}{y_{ba}^{(t)}}(y - y_{ba}^{(t)})^2 \end{aligned} \tag{16}$$

is an auxiliary function for F_{ab}.

Proof. We only need to show that $G(y, y_{ba}^{(t)}) \geq F_{ba}(y)$ because $G(y, y) = F_{ba}(y)$is obvious. There for, we compare the Taylor series expansion of $F_{ba}(y)$

$$\begin{aligned} F_{ba}(y) = F_{ba}(y_{ba}^{(t)}) &+ F'_{ba}(y_{ba}^{(t)})(y - y_{ba}^{(t)}) \\ &+ [(\mathbf{A}^T\mathbf{A}\mathbf{G})_{bb} + \lambda\mathbf{L}_{aa}](y - y_{ba}^{(t)})^2. \end{aligned} \tag{17}$$

$\square$

With Eq. (16) to find that $G(y, y_{ba}^{(t)}) \geq F_{ba}(y)$ is equivalent to

$$\frac{(\mathbf{A}^T\mathbf{A}\mathbf{Y}\mathbf{G})_{ba} + \lambda(\mathbf{Y}\mathbf{D})_{ba}}{y_{ba}^{(t)}} \geq (\mathbf{A}^T\mathbf{A}\mathbf{G})_{bb} + \lambda\mathbf{L}_{aa}. \tag{18}$$

In fact, we have

$$(\mathbf{A}^T\mathbf{A}\mathbf{Y}\mathbf{G})_{ba} = \sum_{l=1}^{k} y_{bl}^{(t)}(\mathbf{A}^T\mathbf{A}\mathbf{G})_{la} \geq (\mathbf{A}^T\mathbf{A}\mathbf{G})_{bb} y_{ba}^{(t)}. \tag{19}$$

And

$$\begin{aligned} \lambda(\mathbf{Y}\mathbf{D})_{\mathrm{ba}} &= \lambda \sum_{j=1}^{m} y_{bj}^{t}\mathbf{D}_{ja} \geq \lambda y_{ba}^{t}\mathbf{D}_{aa} \\ &\geq \lambda y_{ba}^{t}(\mathbf{D} - \mathbf{W})_{aa} = \lambda y_{ba}^{t}\mathbf{L}_{aa}. \end{aligned} \tag{20}$$

Thus Eq. (18) holds and $G(y, y_{ba}^{(t)}) \geq F_{ba}(y)$. Now we can demonstrate the convergence of theorem:

Proof of theorem. Replacing $G(y, y_{ba}^{(t)})$ in Eq. (14) by Eq. (16) results in the following update rules:

$$\begin{aligned} y_{ba}^{(t+1)} &= y_{ba}^{(t)} - y_{ba}^{(t)} \frac{F'_{ba}(y_{ba}^{(t)})}{2(\mathbf{A}^T\mathbf{A}\mathbf{Y}\mathbf{G})_{ba} + 2\lambda(\mathbf{Y}\mathbf{D})_{ba}} \\ &= y_{ba}^{(t)} \frac{(\mathbf{A}^T\mathbf{X}\mathbf{G} + \lambda\mathbf{Y}\mathbf{W})_{ba}}{(\mathbf{A}^T\mathbf{A}\mathbf{Y}\mathbf{G})_{ba} + \lambda(\mathbf{Y}\mathbf{D})_{ba}}. \end{aligned} \tag{21}$$

$\square$

Since Eq. (16) is an auxiliary function, F_{ba} is non-increasing under the update rules.

ACKNOWLEDGMENTS

This work was supported in part by the NSFC under grant Nos. 61572284, 61502272, 61572283, 61370163 and 61272339; the Shandong Provincial Natural Science Foundation, under grant No. BS2014DX004; China Postdoctoral Science Foundation funded project, No.2014M560264; Shenzhen Municipal Science and Technology Innovation Council (Nos. JCYJ20140904154645958, JCYJ20140417172417174, and CXZZ20140904154910774).

REFERENCES

[1] H. B. Saber and M. Elloumi, "DNA microarray data analysis: A new survey on biclustering," *Int. J. Comput. Biol.*, vol. 4, no. 1, pp. 21–37, 2015.

[2] Y. Li and Z. Zhang, "Computational Biology in microRNA," *Wiley Interdisciplinary Rev., RNA*, vol. 6, pp. 435–452, 2015.

[3] B. Bayar, N. Bouaynaya, and R. Shterenberg, "Probabilistic non-negative matrix factorization: Theory and application to microarray data analysis," *J. Bioinformat. Comput. Biol.*, vol. 12, no. 01, pp. 1450001, 2014.

[4] C.-H. Zheng, L. Zhang, T.-Y. Ng, C. K. Shiu, and D.-S. Huang, "Metasample-based sparse representation for tumor classification," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 8, no. 5, pp. 1273–1282, Sep./Oct. 2011.

[5] J.-X. Liu, Y.-L. Gao, Y. Xu, C.-H. Zheng, and J. You, "Differential expression analysis on RNA-Seq count data based on penalized matrix decomposition," *IEEE Trans. NanoBiosci.*, vol. 13, no. 1, pp. 12–18, Mar. 2014.

[6] J.-X. Liu, C.-H. Zheng, and Y. Xu, "Extracting plants core genes responding to abiotic stresses by penalized matrix decomposition," *Comput. Biol. Med.*, vol. 42, no. 5, pp. 582–589, 2012.

[7] J.-X. Liu, Y. Xu, C.-H. Zheng, H. Kong, and Z.-H. Lai, "RPCA-based tumor classification using gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 12, no. 4, pp. 964–970, Jul./Aug. 2015.

[8] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis," *J. Mach. Learning Res.*, vol. 11, pp. 517–553, 2010.

[9] P. K. Yalavarthy, B. W. Pogue, H. Dehghani, and K. D. Paulsen, "Weight-matrix structured regularization provides optimal generalized least-squares estimate in diffuse optical tomography," *Med. Phys.*, vol. 34, no. 6, pp. 2085–2098, 2007.

[10] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learning Res.*, vol. 5, pp. 1457–1469, 2004.

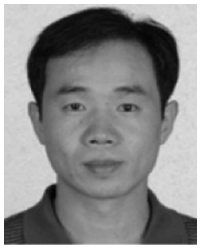
[11] K. Devarajan, "Nonnegative matrix factorization: An analytical and interpretive tool in computational biology," *PLoS Comput. Biol.*, vol. 4, no. 7, pp. e1000029, 2008.

[12] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proc. Natl. Acad. Sci.*, vol. 101, no. 12, pp. 4164–4169, 2004.

[13] R. Giancarlo and F. Utro, "Speeding up the consensus clustering methodology for microarray data analysis," *Algorithms Molecular Biol.*, vol. 6, no. 1, p. 1, 2011.

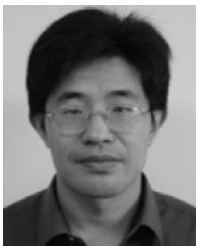
[14] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.
[15] D. Wang, Y.-L. Gao, J.-X. Liu, J.-G. Yu, and C.-G. Wen, "Application of graph regularized non-negative matrix factorization in characteristic gene selection," in *Intelligent Computing Theories and Methodologies*. New York, NY, USA: Springer, 2015, pp. 601–611.
[16] J.-X. Liu, Y. Xu, Y.-L. Gao, C.-H. Zheng, D. Wang, and Q. Zhu, "A class-information-based sparse component analysis method to identify differentially expressed genes on RNA-Seq data," May 2015, (in press).
[17] J. Huang, F. Nie, H. Huang, and C. Ding, "Robust manifold nonnegative matrix factorization," *ACM Trans. Knowl. Discovery Data*, vol. 8, no. 3, pp. 11, 2014.
[18] X. Long, H. Lu, Y. Peng, and W. Li, "Graph regularized discriminative non-negative matrix factorization for face recognition," *Multimedia Tools Appl.*, vol. 72, no. 3, pp. 2679–2699, 2014.
[19] B. Geng, D. Tao, C. Xu, Y. Yang, and X.-S. Hua, "Ensemble manifold regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1227–1233, Jun. 2012.
[20] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint ℓ2, 1-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
[21] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using l21-norm," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 673–682.
[22] S. Yang, C. Hou, C. Zhang, Y. Wu, and S. Weng, "Robust non-negative matrix factorization via joint sparse and graph regularization," in *Proc. Int. Joint Conf. Neural Netw.*, 2013, pp. 1–5.
[23] N. Guan, D. Tao, and Z. Luo, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 2030–2048, Jun. 2011.
[24] J. Liu, J.-X. Liu, Y.-L. Gao, X.-Z. Kong, X.-S. Wang, and D. Wang, "A P-norm robust feature extraction method for identifying differentially expressed genes," *PloS One*, vol. 10, no. 7, pp. e0133124, 2015.
[25] S. Ortega-Martorell, P. J. Lisboa, A. Vellido, M. Julià-Sapé, and C. Arús, "Non-negative matrix factorisation methods for the spectral decomposition of MRS data from human brain tumours," *BMC Bioinformat.*, vol. 13, no. 1, p. 38, 2012.
[26] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, 2015.
[27] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 585–591.
[28] C. Deng, H. Xiaofei, H. Jiawei, and S. H. Thomas, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.
[29] X. Liu, D. Zhai, D. Zhao, G. Zhai, and W. Gao, "Progressive image denoising through hybrid graph Laplacian regularization: A unified framework," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1491–1503, Apr. 2014.
[30] G.-C. Wu and D. Baleanu, "Variational iteration method for the Burgers' flow with fractional derivatives—New Lagrange multipliers," *Appl. Math. Model.*, vol. 37, no. 9, pp. 6183–6190, 2013.
[31] F. Facchinei, C. Kanzow, and S. Sagratella, "Solving quasi-variational inequalities via their KKT conditions," *Math. Program.*, vol. 144, nos. 1/2, pp. 369–412, 2014.
[32] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.
[33] Y. Zhang and Z.-C. Mu, "Ear recognition based on improved NMFSC," *J. Comput. Appl.*, vol. 4, pp. 010, 2006.
[34] G. Nyamundanda, I. C. Gormley, and L. Brennan, "A dynamic probabilistic principal components model for the analysis of longitudinal metabolomics data," *J. Royal Statist. Soc., Ser. C (Appl. Statist.)*, vol. 63, pp. 763–782, 2014.
[35] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.
[36] C.-H. Zheng, D.-S. Huang, L. Zhang, and K. Xiang-Zhen, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 4, pp. 599–607, Jul. 2009.
[37] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," *Bioinformatics*, vol. 21, no. 21, pp. 3970–3975, 2005.
[38] M. A. Shipp, N. Ross Ken, and P. Tamayo, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Med.*, vol. 8, no. 1, pp. 68–74, 2002.
[39] A. Bhattacharjee, W. G. Richards, and J. Staunton, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 24, pp. 13790–13795, 2001.
[40] Y. Shizhun, H. Chenping, and Z. Changshui, "Robust nonnegative matrix factorization via joint sparse and graph regularization for transfer learning," *Neural Comput. Appl.*, vol. 23, no. 2, pp. 541–559, 2013.
[41] M.-Y. Wu, D.-Q. Dai, X.-F. Zhang, and Y. Zhu, "Cancer subtype discovery and biomarker identification via a new robust network clustering algorithm," *PloS One*, vol. 8, no. 6, p. e66256, 2013.
[42] R. Siegel, E. Ward, O. Brawley, and A. Jemal, "Cancer statistics, 2011," *CA: Cancer J. Clin.*, vol. 61, no. 4, pp. 212–236, 2011.
[43] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, and C. Lau, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.
[44] B. J. Druker and M. Talpaz, "Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia," *New Engl. J. Med.*, vol. 344, no. 14, pp. 1031–1037, 2001.
[45] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc. Ser. B (Methodol.)*, vol. 39, pp. 1–38, 1977.

Dong Wang received the BS degree in electronic information science and technology from HeZe University, China, in 2013. She is a master's degree candidate in communication and information system at QuFu Normal University, China. Her research interests include pattern recognition, bioinformatics, and data mining.



Jin-Xing Liu received the BS degree in electronic information and electrical engineering from Shandong University, China, in 1993, the MS degree in control theory and control engineering from QuFu Normal University, China, in 2003, and the PhD degree in computer simulation and control from the South China University of Technology, China, in 2008. He is an associated professor in the School of Information Science and Engineering, Qufu Normal Univeristy. His research interests include pattern recognition, machine learning, and bioinformatics.

Ying-Lian Gao received the BS and MS degrees from QuFu Normal University, China, in 1997 and 2000, respectively. Now, she is currently with the Library at Qufu Normal University. Her current interests include data mining and pattern recognition.



Chun-Hou Zheng received the BS degree in physics education and the MS degree in control theory and control engineering from QuFu Normal University, China, in 1995 and 2001, respectively, and the PhD degree in pattern recognition and intelligent system from the University of Science and Technology of China in 2006. He is currently with the College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui, China. His research interests include pattern recognition and bioinformatics.

Yong Xu received the BS and MS degrees from the Air Force Institute of Meteorology, China, in 1994 and 1997, respectively. He then received the PhD degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology in 2005. From May 2005 to April 2007, he worked at the Shenzhen Graduate School, Harbin Institute of Technology (HIT), as a postdoctoral research fellow. Now, he is a professor at the Shenzhen Graduate School, HIT. His current interests include pattern recognition, machine learning, and bioinformatics.