Tumor Clustering Using Nonnegative Matrix Factorization With Gene Selection

Chun-Hou Zheng, De-Shuang Huang, *Senior Member, IEEE*, Lei Zhang, *Member, IEEE*, and Xiang-Zhen Kong

***Abstract*—Tumor clustering is becoming a powerful method in cancer class discovery. Nonnegative matrix factorization (NMF) has shown advantages over other conventional clustering techniques. Nonetheless, there is still considerable room for improving the performance of NMF. To this end, in this paper, gene selection and explicitly enforcing sparseness are introduced into the factorization process. Particularly, independent component analysis is employed to select a subset of genes so that the effect of irrelevant or noisy genes can be reduced. The NMF and its extensions, sparse NMF and NMF with sparseness constraint, are then used for tumor clustering on the selected genes. A series of elaborate experiments are performed by varying the number of clusters and the number of selected genes to evaluate the cooperation between different gene selection settings and NMF-based clustering. Finally, the experiments on three representative gene expression datasets demonstrated that the proposed scheme can achieve better clustering results.**

***Index Terms*—Clustering, gene expression data, independent component analysis (ICA), nonnegative matrix factorization (NMF), tumor.**

I. Introduction

THE Rapid development of microarray technologies, which can simultaneously assess the expression level of thousands of genes, makes the precise, objective, and systematic analyses, and diagnoses of human cancers possible. A reliable and precise identification of the type of tumors is essential for effective treatment of cancer. Until now, many techniques have been proposed and used to analyze gene expression data, which have demonstrated the potential power for tumor-type identification [1]–[6]. The microarray data typically contain thousands of genes on each chip, and the number of the collected tumor samples is much smaller than that of genes. So it is a typical "large p, small n" problem [7], i.e., the number of predictor variables p is much greater than that of available samples n. The particular condition $p \gg n$ makes most of the standard statistical methods difficult to use from both analytical and interpretative points of view. For example, including too many variables may decrease the accuracy in clustering the samples, and make the cluster rules difficult to set. The inclusion of irrelevant or noisy variables may also degrade the overall performances of the estimated cluster rules.

Despite these difficulties, the clustering and classification methods from the areas of statistical machine learning have been applied to cancer identification using molecular gene expression data [2]–[4], [8], [9]. In this paper, we are interested in unsupervised clustering-based cancer class discovery, instead of supervised classification. Clustering does not require knowing *a priori* the classes of training datasets, which are required by the supervised learning methods. To date, many well-known unsupervised methods, such as hierarchical clustering (HC), self-organizing maps (SOMs), nonnegative matrix factorization (NMF), and its extensions have been used successfully for cancer clustering [2], [4], [5], [10]–[12]. Brunet *et al.* [4] demonstrated that NMF is more accurate than HC and more stable than SOM. Gao and George [12] showed that the results can be improved by using the sparse NMF (SNMF). Kong *et al.* [5] applied the NMF with sparseness constraint (NMFSC) to a microarray dataset, and they concluded that NMFSC performs better than NMF by choosing appropriate degree of sparseness but it does not perform better than SNMF.

Though the aforementioned NMF-based clustering algorithms are useful, one disadvantage of them is that they cluster the microarray dataset with thousands of genes directly, which makes the clustering result not very satisfying. To overcome this problem, in this paper, we propose to perform gene selection before clustering to reduce the effect of irrelevant or noisy variables, so as to achieve a better clustering result.

Gene selection has been used for cell classification [13]–[16]. A comparative study of discrimination methods in the context of cancer classification with filtered sets of genes can be found in [16] and [17]. Feature (gene) selection in such context has the following biological explanation: most of the abnormalities in cell behavior are due to irregular gene activities, and thus, it is critical to highlight these particular genes. Daniela *et al.* [13] employed independent component (IC) analysis (ICA) [18] to select subsets of genes that might be relevant to different tumors. Afterwards, they applied the supervised method to the selected

Manuscript received February 20, 2008; revised October 6, 2008 and February 11, 2009. First published April 14, 2009; current version published July 6, 2009. This work was supported by the National Science Foundation of China under Grant 30700161, by the National Fundamental Research Program of China (973 Program) under Grant 2007CB311002, by the National High Technology Research and Development Program of China (863 Program) under Grant 2007AA01Z167, by the Guide Project of Innovative Base of Chinese Academy of Sciences (CAS) under Grant KSCX1-YW-R-30, by the China Postdoctoral Science Foundation under Grant 20070410223, and by the Encouragement Foundation for Young and Middleaged Scientist of Shandong Province under Grant 2008BS01010.

C.-H. Zheng was with the Intelligent Computing Laboratory, Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China. He is now with the College of Information and Communication Technology, Qufu Normal University, Rizhao 276826, China (e-mail: zhengch99@126.com).

D.-S. Huang is with the Intelligent Computing Laboratory, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Anhui 230031, China (e-mail: dshuang@iim.ac.cn).

L. Zhang is with the Biometric Research Center, Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: cslzhang@comp.polyu.edu.hk).

X.-Z. Kong is with the College of Information and Communication Technology, Qufu Normal University, Rizhao 276826, China (e-mail: liumickey@126.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2009.2018115

gene expression data for cell classification, and showed that the gene selection strategy is efficient and feasible.

The successful use of ICA and NMF in processing gene expression data [4], [12], [13], [19], [20] inspires us to combine them for improving the clustering performance. In this paper, we first employed ICA to select a subset of genes, and then used NMF and its extensions to cluster the tumors on the subset of genes selected by ICA. During gene selection, since gene expression profiles are typically super-Gaussian, we should exploit not only the second-order statistics of the data structure but also the higher order statistics. Therefore, ICA is employed to extract the statistically independent features. By a series of elaborate experiments, the suitable number of ICs and the number of selected genes are analyzed, and the number of genes that yields the best clustering stability can be determined. To find out the result of which clustering algorithm cooperating with the proposed gene selection method is the best, we compared the results by using NMF, SNMF, and NMFSC, respectively, on three different gene expression datasets.

The rest of the paper is organized as follows. Section II describes the ICA model of gene expression data and the ICA based gene selection method. Section III presents the NMF algorithm and the principles of clustering using NMF. The tumor clustering experiments are performed in Section IV. Section V concludes the paper and outlines directions of future work.

II. Gene Selection by ICA

ICA can be regarded as a dimension reduction technique, which decomposes the input multivariate dataset into statistically independent components (ICs). ICA can reduce the effects of noise or artifacts on the signal and is efficient for separating mixed signals [18], [21]. Recently, more and more successful applications of ICA into microarray data analysis were reported to extract expression modes of genes [19], [20], [22], [23]. This section will present the ICA model of gene expression data and the gene selection method based on ICA.

A. Independent Component Analysis

ICA is a useful extension to principal component analysis (PCA), which was originally developed for blind separation of independent sources from their linear mixtures [18]. It has been used in various applications of auditory signal separation, medical signal processing, and so on. Unlike PCA, where the aim is to decorrelate the dataset, ICA aims to make the transformed coefficients mutually independent (or as independent as possible). This implies that the higher order dependencies will be removed by the ICA expansion.

Considering a $p \times n$ data matrix X, whose columns c_j $(j = 1, \ldots, n)$ represent the observational variables, the ICA model of X can be written as (in some ICA literature, the problem is formulated by using the transposed matrix X^T)

$$X = SA \tag{1}$$

where S is a $p \times n$ source matrix and A an $n \times n$ mixing matrix. Vectors s_q, the columns of S, are assumed to be statistically independent and are called as the ICs of S. Model (1) implies

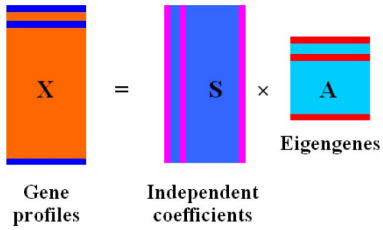


Fig. 1. ICA model of gene expression data. Each gene profile in the data matrix is considered to be a linear combination of underlying basis expression profiles (eigengenes) in the matrix A (the rows in A). Each of the basis expression profiles is associated with a set of independent "causes (coefficients)," given by a vector of coefficients in S. The basis profiles are estimated by $A = W^{-1}$, where W is the learned ICA weight matrix.

that the columns of X are linear mixtures of the ICs. The statistical independence between s_q can be measured by using mutual information $I = \sum_q H(s_q) - H(S)$, where $H(s_q)$ is the marginal entropy of the variable s_q and $H(S)$ the joint entropy of S. Estimating the ICs can be accomplished by finding the right linear combinations of the observational variables. We can invert the mixing matrix such that

$$S = XA^{-1} = XW. \tag{2}$$

Then, an ICA algorithm is used to find a projection matrix W such that the columns of S are as statistically independent as possible.

Several algorithms have been proposed to implement ICA, such as FastICA [24] and JADE [25], etc. In this paper, we employ the FastICA algorithm to model the gene expression data, considering its efficiency in processing large-scale dataset. The FastICA has been widely used to process gene expression data [13], [19], [22], [26]. In FastICA, the mutual information is approximated by

$$J(s_q) = (E\{G(s_q)\} - E\{G(\varsigma)\})^2 \tag{3}$$

where G is an arbitrary nonquadratic function and ς a Gaussian distributed variable. The interested readers can refer to literature [24] for details. ICA can remove linear correlations as well as higher order dependencies in the data. It also allows some flexibility in scaling and sorting by convention. The ICs are generally scaled to unit deviation, while their signs and orders can be chosen arbitrarily.

B. ICA Models of Gene Expression Data

The gene expression dataset can be represented by a $p \times n$ matrix X $(p \gg n)$, whose element x_{ij} is the expression level of the *i*th gene in the *j*th assay $(1 \le i \le p, 1 \le j \le n)$. The *n*-dimensional vector r_i, i.e., the *i*th row of X, denotes the expression profile of the *i*th gene. Alternatively, the *p*-dimensional vector c_j, i.e., the *j*th column of X, is the snapshot of the *j*th assay (cell sample). We suppose that the dataset has been preprocessed and normalized, i.e., every cell sample has zero mean and unit standard deviation.

In the ICA model for gene expression data, the linear ICA model $X = SA$ represents the gene expression profiles by a new set of basis vectors (the rows of A, as shown in Fig. 1). This

idea comes from the following assumptions. First, the gene expression profiles are determined by a combination of hidden variables, which are called "expression modes (eigengenes)." Second, the genes' responses to these variables can be approximated by linear functions [16], [27]. Expression profile q (the qth row of X) is characterized by all of the eigengenes (all the rows of A) and by its linear independent influences on the eigengenes (the qth row of S). In this paper, we use this idea to find a good set of basis profiles to represent gene expression data so that the subsets of genes relevant to cell classification can be selected.

C. Gene Selection: An ICA-based Solution

One of the objectives of this paper is to propose a method to select subsets of genes that could be relevant for cell clustering. The selection is performed by projecting the genes onto the desired directions obtained by ICA. In particular, the distribution of gene expression levels on a cell is "approximately sparse," with heavy tails and a pronounced peak in the middle. Due to this, the projections obtained by ICA should emphasize this sparseness. Highly induced or repressed genes, which may be useful in cell clustering, should lie in the tails of the distributions of $s_j (j = 1, \mathrm{K}, z)$, where z is the actual number of ICs we estimated from the gene expression data. Since these directions are independent, they may catch different aspects of the data structure that could be useful for classification tasks [13]. The proposed gene selection method is based on a ranking of the p genes. This ranking process is introduced as follows.

Step 1: z-ICs $s_1, \ldots, s_z$ with zero mean and unit variance are extracted from the gene expression dataset using ICA.

Step 2: For gene $l (l = 1, \ldots, p)$, the absolute score on each component $|s_{lj}|$ is computed. These z scores are synthesized by retaining the maximum one, denoted by $g_l = \max_j |s_{lj}|$.

Step 3: The p genes are sorted in increasing order according to the maximum absolute scores $\{g_1, \ldots, g_p\}$, and for each gene, the rank $r(l)$ is computed.

In our experiments, we found that ICA is not always reproducible when used to analyze gene expression data. This problem has also been reported in [19]. It is because that the ICA algorithm may converge to local optima [20]. To solve this problem, we run the independent source estimation process for 100 times with different random initializations. In each time, we chose the subset of the last *m* genes (with $m \leq p$). The rationale behind this is that these *m* genes show, with respect to at least one of the components, a behavior across the cells that differs most from that of the bulk of the genes [13]. After running the algorithm 100 times, we obtained $100 \times m$ genes, of which most of them are reduplicate. The selected genes were then sorted in descending order according to the selected frequency of each gene. Finally, we selected the first m genes for the next step of clustering.

Obviously, the proposed strategy has two key parameters, i.e., the number of ICs z and the number of selected genes m. We will describe how to determine them in Section IV-A.

III. Clustering With NMF

NMF can reduce the dimensionality of expression data from thousands of gene to metagenes. Coupled with a model selection mechanism, NMF can be an efficient method to identify distinct molecular patterns and a powerful tool for class discovery. Brunet *et al.* [4] demonstrated the ability of NMF to recover meaningful biological information from cancer-related microarray data. NMF also appears to have advantages over other methods, such as HC and SOMs, because HC imposes a stringent tree structure on the data, and is highly sensitive to the metric used to assess similarity, and SOM can be unstable, yielding different decompositions of the data with different initial conditions. However, standard NMF cannot control the sparseness of the decomposition, and thus, does not always yield a parts-based representation. Some research groups have proposed to impose sparseness constraints on NMF [27]–[30]. It has been shown that the extensions of NMF, e.g., SNMF [12] and NMFSC [5], coupled with the model selection mechanism [4], could improve cancer class discovery on the microarray datasets [7], [12].

A. NMF Algorithm

We now represent the selected gene expression data as a matrix Y of size $m \times n$, whose rows contain the expression levels of the m selected genes in the n cell samples, and each column represents the expression level of all genes in one sample. All the entries in the gene expression matrix are nonnegative. The NMF methods resort to factor the gene expression matrix Y into the product of two matrices of nonnegative entries

$$Y \approx VH \tag{4}$$

where matrix V is of size $m \times k$ with each of the k columns defining a metagene, matrix H is of size $k \times n$ with each of the n columns representing the metagene expression pattern of the corresponding sample, and k is a desired rank. The method starts by randomly initializing matrices V and H, which are iteratively updated to minimize a divergence function. The function is related to the Poisson likelihood of generating Y from V and H

$$D = \sum_{ij} Y_{ij} \log \left(\frac{Y_{ij}}{(VH)_{ij}} \right) - Y_{ij} + (VH)_{ij}. \tag{5}$$

At each step, V and H are updated by using the coupled divergence equations [31]

$$H_{au} \leftarrow H_{au} \frac{\sum_i V_{ia} Y_{iu} / (VH)_{iu}}{\sum_k V_{ka}} \tag{6}$$

$$V_{ia} \leftarrow V_{ia} \frac{\sum_u H_{au} Y_{iu} / (VH)_{iu}}{\sum_v H_{av}}. \tag{7}$$

B. Clustering Using NMF

In NMF model, each entry v_{ij} in V is the coefficient of gene i in metagene j and each entry h_{ij} in H represents the expression level of metagene i in sample j. In such a factorization, matrix H can be used to group the n samples into k clusters [4]. Each cell sample is placed into a cluster corresponding to the most

highly expressed metagene in the sample in [4], i.e., sample j is placed in cluster i if h_{ij} is the largest entry in column j.

Although NMF has been successfully used in several applications, it does not always result in parts-based representations. To solve this problem, Hoyer [29] extended the NMF framework by including an adjustable sparseness parameter. SNMF and NMFSC are extensions to those ideas [27], [29], [30]. The main improvement in them is that the sparseness can be adjusted explicitly, rather than implicitly. These algorithms were used for tumor class clustering in [5] and [12]. The detailed algorithm of SNMF and NMFSC can be found in [12] and [27]. In this paper, except for selecting genes using ICA, we will also study the influence of the sparseness of the factors V and H.

On the other hand, when using NMF to group the samples into clusters, there are several problems that need to be resolved. Among them, one key issue is which k can decompose the samples into "meaningful" clusters. Another problem is that the NMF algorithm may or may not converge to the same solution in each run with the random initial conditions. So, how to evaluate the stability of clustering associated with a given rank k? This is still an open problem.

The authors in [4] developed a nice model selection method based on consensus clustering [32]. The basic idea is that if a clustering to k classes is strong, the cluster assignment of samples should not vary much from random starting points. After running with many different random initial points, a consensus matrix $\bar{C}$ is computed to evaluate the stability of clustering associated with the given k. The entries of $\bar{C}$ range from 0 to 1 and reflect the probability that each pair of samples cluster together. If a clustering is stable, the entries of $\bar{C}$ will be close to 0 or 1. The dispersion between 0 and 1 thus measures the reproducibility of the class assignments with respect to the random initial conditions. A reordered matrix of $\bar{C}$ can be used for visual inspection, which can serve as similarity measure among samples (refer to Fig. 2(a) for the details). Quantitatively, the stability for each value of k can be measured through the cophenetic correlation coefficient $\rho_k(\bar{C})$, which indicates the dispersion of the consensus matrix $\bar{C}$ [4].

In a perfect consensus matrix (all entries $= 0$ or 1), the cophenetic correlation coefficient $\rho_k = 1$. When the entries are scattered between 0 and 1, the cophenetic correlation coefficient is less than 1. Roughly speaking, the more stable the cluster assignment is, the greater the coefficient ρ_k is. Therefore, by observing how ρ_k evolves as k increases, we can select the value of k when the magnitude of the coefficient starts to fall. Interested readers may refer to [33] and [34] for more details about how the coefficient ρ_k is calculated. For example, in Fig. 2(b), we can see that ρ_k drops when k increases from 2 to 5. This implies that a two-cluster split of the samples is more stable than others.

IV. Experimental Results

To evaluate the performance of the proposed method, we apply it to three publicly available datasets, i.e., leukemia dataset [2], central nervous system embryonal tumors dataset [9], and medulloblastoma dataset [9]. We first employed ICA to select m genes, and then applied NMF and its extensions, i.e., SNMF and NMFSC, to cluster the tumors using the selected genes. To demonstrate the efficiency of the proposed strategy, we applied NMF, SNMF, and NMFSC, respectively, to the subsets of selected genes for cancer class discovery. The clustering accuracy is measured by the following formula [35]:

$$AC = \frac{\sum_{i=1}^{n} I(j_i)}{n} \tag{8}$$

where $I(j_i)$ is 1 if the cluster assignment is correct for sample j_i and 0 if the cluster assignment is incorrect. The clustering accuracy is computed according to the well-known classification label of the tumor dataset [2], [9]. Another criterion for demonstrating the efficiency of the proposed strategy is the cophenetic correlation coefficient $\rho_k(\bar{C})$, which can measure the clustering stability.

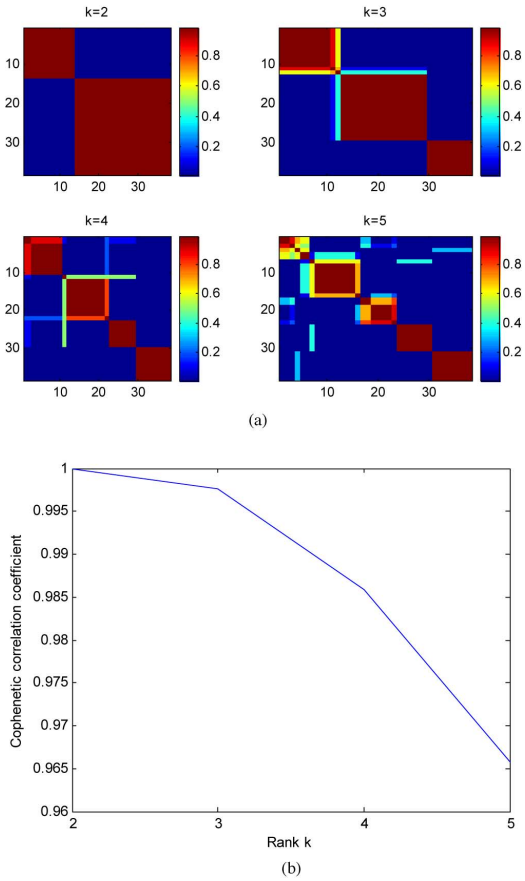


Fig. 2. (a) Reordered consensus matrices for ranks 2–5 of NMF using the leukemia dataset of 38 bone marrow samples with 5000 most highly varying genes. Deep blue color corresponds to a numerical value of 0 and means that the samples are never assigned to the same cluster. Dark red color corresponds to 1 and means that the samples always appear in the same cluster. The 0–1 pattern indicates highly robust classification. (b) Corresponding cophenetic correlation coefficients for hierarchically clustered matrices in (a). ρ drops when k increases from 2 to 5, indicating a two-cluster split of the samples is more stable than others.

A. Leukemia Dataset

The leukemia dataset has become a benchmark in cancer classification. In this dataset, the distinction between acute myelogenous leukemia (AML) and acute lymphoblastic leukemia

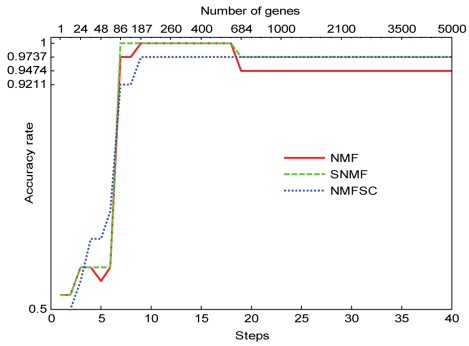


Fig. 3. Clustering accuracy rate for leukemia dataset, where $z = 6$ for ICA, and $k = 2$ for NMF and its extensions, and the values at the top of the figure indicate the number of genes retained at each step.

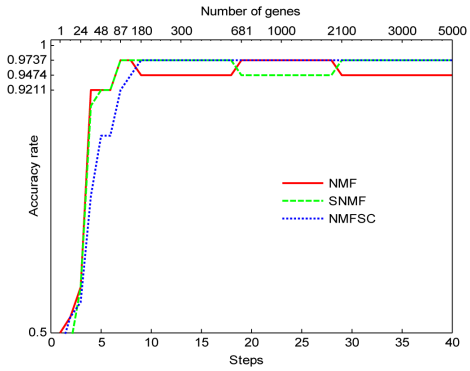


Fig. 4. Clustering accuracy rate for leukemia dataset, where $z = 6$ for ICA, and $k = 3$ for NMF and its extensions, and the values at the top of the figure indicate the number of genes retained at each step.

(ALL), as well as the division of ALL into T and B cell subtypes, is known. The dataset contains $p = 5000$ genes in 38 cells, and consists of 19 cases of B cell ALL (ALL_B), 8 cases of T cell ALL (ALL_T), and 11 cases of AML.

In our method, three parameters, i.e., k, z, and m, need to be determined. We first computed the cophenetic correlation coefficient $\rho_k(\bar{C})$ using the original data, through which the stability for each value of rank k can be measured. In the experiment, we found that ρ drops when the rank increases from 2 to 5 (as shown in Fig. 2(b) for NMF), which indicates that a two-cluster and three-cluster splits of the samples should be more stable than others, i.e., $k = 2$ or $k = 3$.

How to determine the IC number z is still an open problem in ICA. In this paper, we selected z experimentally. Particularly, we run the gene selection procedure with z increasing from 1 to 10. For each value of z, the gene ranking is computed and a sequence of gene subset is selected for five different values of m ranging from 1 to p, e.g., $m = 300$, 500, 800, 1200, and 1500. Then, we cluster the tumors using the selected genes and calculated the cophenetic correlation coefficient $\rho_k(\bar{C})$. The value of z is determined such that it can achieve relatively high $\rho_k(\bar{C})$ values for all the experiments.

In this experiment, we selected $z = 6$ by the aforementioned process. Fig. 3 shows the results obtained with $z = 6$, $k = 2$, and 40 different values of m. Fig. 4 is the result for k= 3. For comparison, we also list the results of [4] and [12] in Figs. 3 and 4, i.e., the accuracies of NMF and SNMF with $m = 5000$. From the two figures, we can see that no matter the rank k is 2 or 3, gene selection by suitable projections can achieve commensurate or better performance than the methods without gene selection, i.e., $m = 5000$.

With the rank $k = 2$, SNMF achieves 100% accuracy rate when the number of retained genes is 86; for NMF, it can also achieve100% accuracy rate with the number of key genes being 183; for NMFSC ($s_v = 0.1$, $s_h = 0$), it could achieve the accuracy rate by 97.37% with 198 genes (misclassified AML_13 to ALL). With rank $k = 3$, the accuracy rates are not improved when compared with those in [4] and [12], yet the proposed algorithm could also achieve 97.37% accuracy rate (AML_13 sample was misclassified to ALL_B) with less key genes. On this dataset, the ICA-based genes selection can make the class structure more evident, i.e., the use of suitable subsets of genes instead of the whole set could yield better clustering performance.

However, since clustering is a type of unsupervised learning, the accuracy rates could not be used as the criterion for selecting the gene number m. We experimentally found that the clustering result is stable when the number of selected genes is changing in a wide range. In practice, the criterion for determining m is as follows. Once k and z are determined, we choose a small number of genes, e.g., $m = 500$, that corresponds to the biggest value of $\rho_k(\bar{C})$ for clustering.

Fig. 5 demonstrates the reordered consensus matrices for $k = 2$ and $k = 3$ when using the leukemia dataset of 38 bone marrow samples with the original 5000 genes as well as the selected genes that have the best stability. The corresponding accuracy rates can be found in Figs. 3 and 4. From Fig. 5, we can see that for NMF and its extensions, when the rank $k = 2$ [see Fig. 5(a)], we can obtain the same clustering stability with the selected genes, yet when $k = 3$ [see Fig. 5(b)], the clustering stability is obviously improved by using the proposed gene selection strategy, especially for NMF and SNMF. In summary, from Figs. 3–5, we can conclude that our method can improve either the accuracy rates or the clustering stability.

To better demonstrate the advantage of the proposed ICA-based gene selection, we also used the genes with high variances for clustering. The experimental results are shown in Figs. 6–8. Comparing these three figures with Figs. 3–5, respectively, we can clearly see that ICA-based gene selection is more effective than the variance-based gene selection method.

B. Central Nervous System Tumors

This dataset is composed of four types of central nervous system embryonal tumors [9]. The dataset used in our experiment contains $p = 5560$ genes in 34 samples representing four distinct morphologies: ten classic medulloblasomas, ten malignant gliomas, ten rhabdoids, and four normals. The previous studies [4] without gene selection, i.e., NMF and SNMF, suggest a four-cluster split with high cophenetic coefficient. The NMF method has two misclassifications. One is assigning a glioma (Brain_MGlio_8) to rhabdoid and the other one is more serious: it incorrectly assigns a rhabdoid sample (Brain_Rhab-10) to normal. Such an assignment will definitely delay the treatment

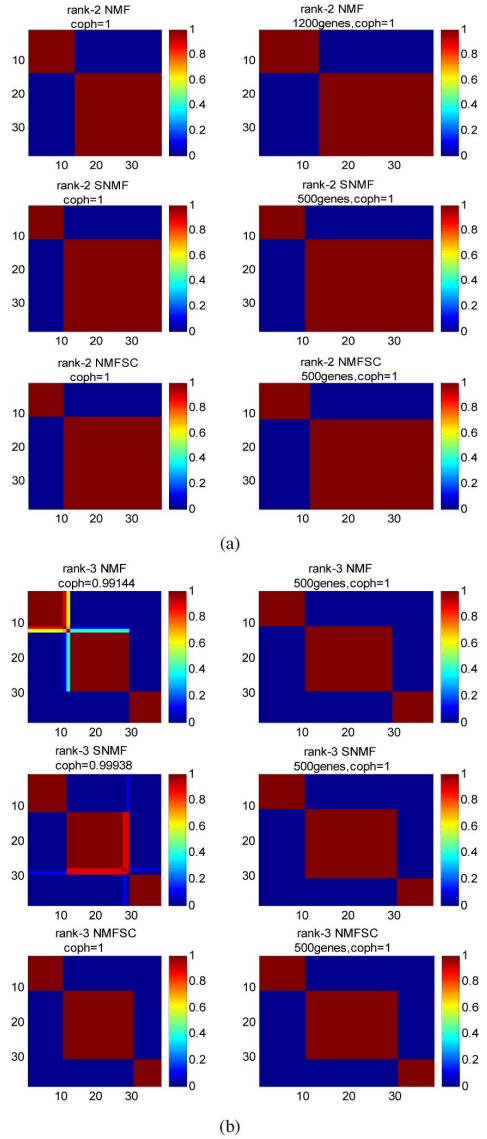


Fig. 5. Reordered consensus matrices and the corresponding cophenetic correlation coefficients for ranks 2–3 of NMF and its extensions, using the leukemia dataset of 38 bone marrow samples (left) with 5000 most highly varying genes and (right) with the selected genes that have the best stability ($z = 6$ for ICA).

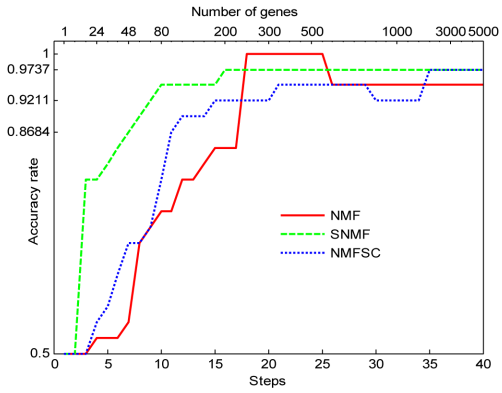


Fig. 6. Clustering accuracy rate for leukemia dataset, where rank $k = 2$ for NMF and its extensions. The genes are ordered based on variance, and the values at the top of the figure indicate the number of genes retained at each step.

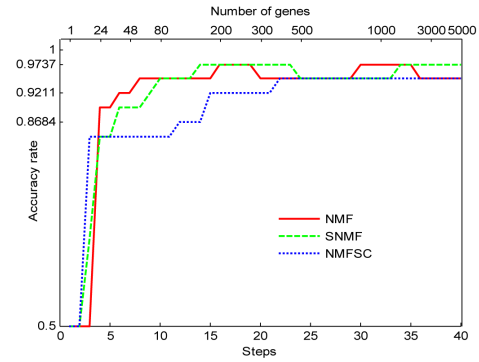


Fig. 7. Clustering accuracy rate for leukemia dataset, where rank $k = 3$ for NMF and its extensions. The genes are ordered based on variance, and the values at the top of the figure indicate the number of genes retained at each step.

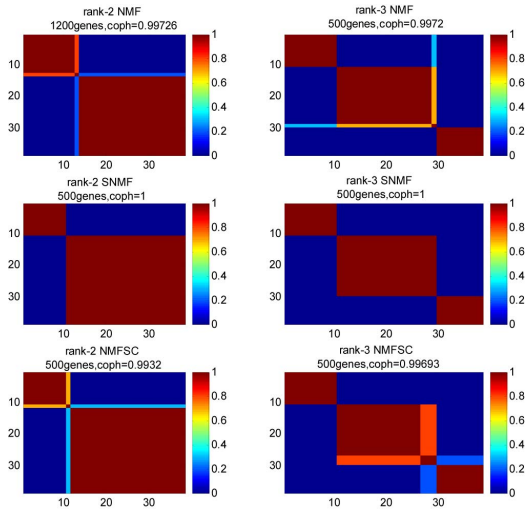


Fig. 8. Reordered consensus matrices and the corresponding cophenetic correlation coefficients for ranks 2–3 of NMF and its extensions, using the ordered genes based on variance for the leukemia dataset.

of the patient, and is thus highly undesired. The SNMF method correctly split the samples into four clusters with high cophenetic coefficient [12]. It has only one misclassification: the same glioma (Brain_MGlio_8) is assigned to rhabdoid. Moreover, it can correctly cluster the four normal samples into a distinct group.

In our study, we adopted the four-cluster suggestion as in [12], i.e., rank $k = 4$ for the NMF and its extensions. We applied NMF and its extensions to the appropriately selected key genes ($m = 1000$) as in the first experiment. The misclassification is the same as that in [12], i.e., the glioma (Brian_MGlio_8) is assigned to rhabdoid. Fig. 9 shows the results for $z = 7$ and 40 different values of m. We also compared the clustering stability with and without gene selection in Fig. 10. One can see that through gene selection, the stability is improved for NMF and SNMF. For NMFSC, the stability is the same as that without gene selection. It can be concluded that for this dataset, both sparseness and gene selection could improve the clustering result. Similar to that in Section IV-A, in this experiment, we also found that ICA-based gene selection is much better than variance-based gene selection.

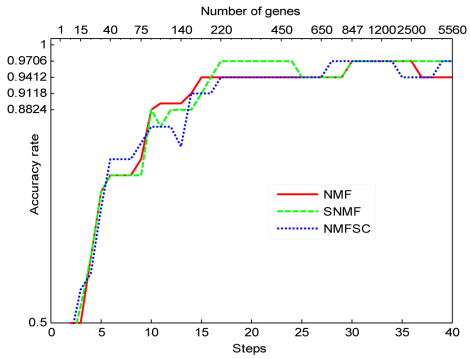


Fig. 9. Clustering accuracy rate for central nervous system tumors dataset, where $z = 7$ for ICA, and $k = 3$ for NMF and its extensions, and the values at the top of the figure indicate the number of genes retained at each step.

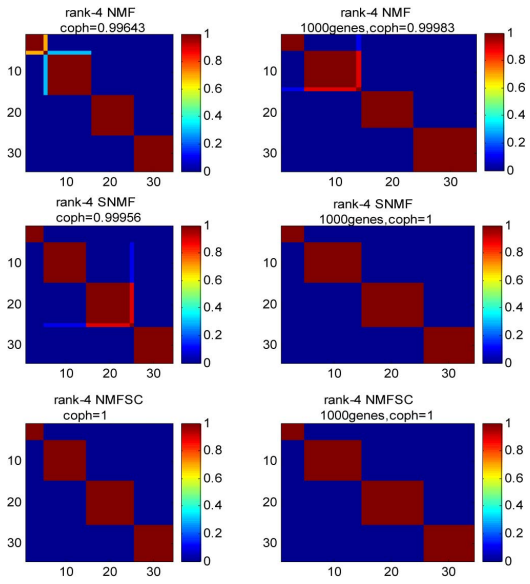


Fig. 10. Reordered consensus matrices and the corresponding cophenetic correlation coefficient for rank 4 of NMF and its extensions using the central nervous system tumors dataset of 34 samples with (left) 5560 genes and (right) the selected genes using ICA ($z = 7$).

C. Medulloblastoma Dataset

We also applied the proposed method to the medulloblastoma dataset [9], which is about childhood brain tumors. The pathogenesis of these tumors is not well understood, but it is generally accepted that there are two known histological subclasses: classic and desmoplastic, whose differences can be clearly seen under the microscope [4]. In our experiment, the dataset contains $p = 5893$ genes, 34 samples. The samples can be divided into 25 classic and nine desmoplasitic medulloblastomas. However, such diagnosis is highly subjective [9], [12].

It has been shown that the straightforward use of basic NMF is better in this case for $k = 2$, 3 and 5 [4], [12]. In our experiment, we set the number of ICs as $z = 8$. Given the fact that the pathogenesis of these tumors is not well understood and desmoplastic medulloblastoma diagnosis is highly subjective, it may raise doubt about the sample labeling [12], i.e., the labels of the samples may be incorrect. Therefore, in this experiment, we did not calculate the accuracy for this dataset. We only showed the reordered consensus matrices m in Fig. 11. The class assignment for medulloblastoma discovered by the clustering with genes selection is listed in Table I. Refer to literatures [4], [5], and [12] for the details of the clustering results without gene selection.

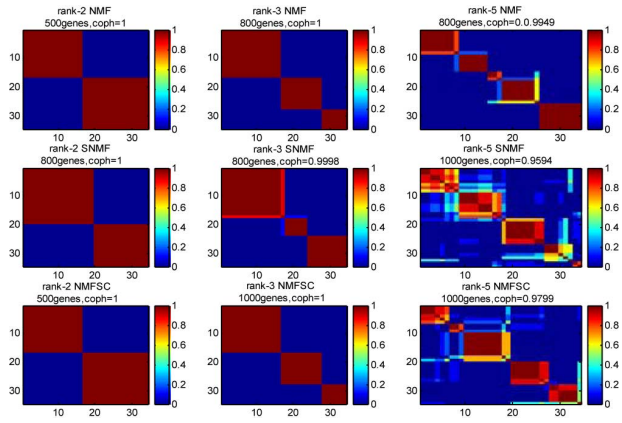


Fig. 11. Reordered consensus matrices and the corresponding cophenetic correlation coefficient for ranks 2, 3, and 5 of NMF and its extensions using the medulloblastoma dataset of 34 samples with 5893 most highly varying genes and the selected genes ($z = 8$ for ICA and $S_v = 0.6$ and $S_h = 0$ for NMFSC).

TABLE I
CLASS ASSIGNMENT FOR MEDULLOBLASTOMA DISCOVERED BY THE CLUSTERING ALGORITHM WITH GENES SELECTION

Sample	Class	NMF			SNMF			NMFSC ($S_v = 0.6, S_h = 0$)		
		k=2 m=500	k=3 m=800	k=5 m=800	k=2 m=800	k=3 m=800	k=5 m=1000	k=2 m=500	k=3 m=1000	k=5 m=1000
Brain_MD_7	1*	2	1	4	2	3	2	1	1	4
Brain_MD_59	1	1	2	3	1	1	5	1	2	5
Brain_MD_20	1	1	1	4	1	3	2	1	1	4
Brain_MD_21	1	1	1	1	1	3	1	1	1	3
Brain_MD_50	1	1	1	4	1	3	2	1	1	2
Brain_MD_49	1	2	3	5	2	2	4	2	3	1
Brain_MD_45	1	1	1	4	1	3	2	1	1	4
Brain_MD_43	1	1	1	4	1	3	2	1	1	4
Brain_MD_8	1	1	1	4	1	3	2	1	1	4
Brain_MD_42	1	2	3	2	2	2	3	2	3	2
Brain_MD_1	1	2	3	1	2	2	3	2	3	1
Brain_MD_4	1	2	3	2	2	2	3	2	3	1
Brain_MD_55	1	2	3	2	2	2	3	2	3	1
Brain_MD_41	1	1	1	4	1	3	2	1	1	4
Brain_MD_37	1	1	2	3	1	1	5	1	2	5
Brain_MD_3	1	2	3	2	2	2	3	2	3	1
Brain_MD_34	1	2	3	2	2	2	3	2	3	3
Brain_MD_29	1	1	1	4	1	3	2	1	1	4
Brain_MD_13	1	2	3	2	2	2	3	2	3	1
Brain_MD_24	1	2	3	2	2	2	3	2	3	1
Brain_MD_65	1	1	1	4	1	3	2	1	1	4
Brain_MD_5	1	1	2	3	1	1	5	1	2	5
Brain_MD_66	1	1	2	3	1	1	5	1	2	5
Brain_MD_67	1	1	2	3	1	1	5	1	2	4
Brain_MD_58	1	2	3	2	2	2	3	2	3	1
Brain_MD_53	2#	2	3	5	2	2	4	2	3	1
Brain_MD_56	2	2	3	5	2	2	4	2	3	2
Brain_MD_16	2	2	3	5	2	2	4	2	3	1
Brain_MD_40	2	2	2	5	2	2	4	2	2	5
Brain_MD_35	2	2	3	5	2	2	4	2	3	2
Brain_MD_30	2	2	3	5	2	2	4	2	3	2
Brain_MD_23	2	2	3	5	2	2	4	2	3	2
Brain_MD_28	2	1	1	1	2	3	1	1	1	3
Brain_MD_60	2	1	2	3	1	1	4	1	2	5

1* Classic, 2# Desmoplastic

D. Software and Discussions

The codes for NMF and its extensions can be obtained from http://www.cs.helsinki.fi/u/phoyer/software.html; the codes for FastICA can be obtained from: http://www.cis.hut.fi/projects/ica/fastica/. The MATLAB codes for this paper are available on http://www.comp.polyu.edu.hk/∼cslzhang/NMF_GS_ICA.htm.

It was reported in previous literatures [36], [37] that higher classification accuracy can be achieved by using only a small amount of genes (about 15). However, from our experiments, it can be found that such a small amount of genes cannot achieve high cluster accuracy (see Figs. 3, 4, 6, and 7). In addition, we found that an amount of about 200 genes may achieve higher accuracy but the clustering result is not stable. Nonetheless, from a viewpoint of pathology, cancer may involve a certain amount of genes, e.g., 500 genes, as in our experiments.

So far, many gene selection methods have been proposed [13]– [15], [17], [22]. The reason that we used ICA for gene selection can be explained as follows. First, there is a sound biological interpretation of ICA model for gene expression data [19], [23]. Second, ICA-based gene selection does not need the labels of samples, so it is very suitable for clustering and can be helpful to other clustering methods. Third, ICA has been proven to be an effective gene selection method for tumor classification [13].

Apart from the NMF clustering algorithm, the proposed ICA-based gene selection method can also be coupled with other clustering algorithms such as HC [4] and SOMs [11]. Our experimental results by coupling ICA-based gene selection with HC and SOM clustering also validate its efficiency. Due to the limitation of space, we have not arranged the results in this paper. The reason that we used NMF for clustering is that NMF is more accurate than HC and is more stable than SOM, as indicated by Brunet *et al.* [4] and validated by our experimental results.

V. Conclusion

In this paper, we employed ICA to model the gene expression data for gene selection, and then applied NMF and its extensions, i.e., SNMF and NMFSC to cancer clustering using the selected genes. The proposed method was validated on the leukemia dataset, embryonal tumors dataset from the central nervous system, and the medulloblastoma dataset. It can be found that improved clustering results were achieved by selecting the key genes using ICA. From the experimental results, we can see that the ICA-based gene selection is useful to detect the subsets of relevant genes for tumor clustering, especially when coupled with the NMF clustering method. It should be noted that although the three datasets used in our experiments have similar number of genes, i.e., about 5000, our method has no constraints on the number of genes contained in the data. In fact, our proposed method can be applied to the datasets that have much more genes.

In future, systematic studies on larger datasets will be conducted for more convincing arguments. First, the definition of ICA implies that there is no order of the ICs. In other words, all the z-estimated ICs are assumed to be equally important in the proposed scheme. It is possible, however, to sort these ICs. Hyvärinen *et al.* [38] suggested an ordering criterion using the norm of the columns of the mixing matrix or the value of suitable non-Gaussianity measures on the estimated ICs. This criterion might be adopted to weigh each IC during the construction of gene ranking (for example, by increasing the importance of the most non-Gaussian ones). Second, the issues concerning the selection of number z and number m should be further examined. Finally, in our study, the clustering results may be different when the proposed gene selection method is coupled with different clustering algorithms. In addition, the better results may be achieved by other gene selection methods. Therefore, the interaction between different gene selection methods and other clustering algorithms should be further explored.

References

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Nat. Acad. Sci. USA*, vol. 96, pp. 6745–6750, 1999.

[2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.

[3] M. Bittner, P. S. J. Meltzer, Y. D. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. L. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. C. A. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, and J. Trent, "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, pp. 536–540, 2000.

[4] J.-P. Brunet, P. Tamayo, T. R. Golun, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 12, pp. 4164–416, 2004.

[5] X. Z. Kong, C. H. Zheng, Y. Q. Wu, and L. Shang, "Molecular cancer class discovery using non-negative matrix factorization with sparseness constraint," in *Proc. Int. Conf. Intell. Comput.*, LNCS 2007, vol. 4681, pp. 792–802.

[6] Q. Zhu, H. Cui, K. Cao, and W. C. Chan, "Algorithmic fusion of gene expression profiling for diffuse large B-cell lymphoma outcome prediction," *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 2, pp. 79–88, Jun. 2004.

[7] M. West, "Bayesian factor regression models in the "Large *p*, Small *n*" paradigm," *Bayesian Stat.*, vol. 7, pp. 723–732, 2003.

[8] K. Bryan, P. Cunningham, and N. Bolshakova, "Application of simulated annealing to the biclustering of gene expression data," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 3, pp. 519–525, Jul. 2006.

[9] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. Mclaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, pp. 436–442, 2002.

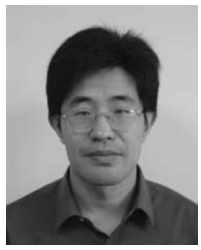
[10] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Nat. Acad. Sci. USA*, vol. 95, pp. 14863–14868, 1998.

[11] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proc. Nat. Acad. Sci. USA*, vol. 96, pp. 2907–2912, 1999.

[12] Y. Gao and C. George, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," *Bioinformatics*, vol. 21, pp. 3970–3975, 2005.

[13] G. C. Daniela, G. Giuliano, P. Marilena, and V. Cinzia, "Variable selection in cell classification problems: A strategy based on independent component analysis," in *Studies in Classification, Data Analysis,*

and Knowledge Organization, Part I. Berlin/Heidelberg, Germany: Springer-Verlag, 2006, pp. 21–29.
[14] Y. Tang, Y. Zhang, and Z. Huang, "Development of two-stage SVM–RFE gene selection strategy for microarray expression data analysis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 4, no. 3, pp. 365–381, Jul.–Sep. 2007.
[15] C. Furlanello, M. Serafini, S. Merler, and G. Jurman, "Entropy-based gene ranking without selection bias for the predictive classification of microarray data," *BMC Bioinf.*, vol. 4, p. 54, 2003.
[16] W. Pan, "A comparative review of statistical methods for discovering differently expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, pp. 546–554, 2002.
[17] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumor using gene expression data," *J. Amer. Stat. Assoc.*, vol. 97, pp. 77–87, 2002.
[18] A. Hyvärinen and E. Oja, "Independent component analysis: algorithm and applications," *Neural Netw.*, vol. 2000, no. 13, pp. 411–430, 2000.
[19] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics*, vol. 18, pp. 51–60, 2002.
[20] P. Chiappetta, M. C. Roubaud, and B. Torrésani, "Blind source separation and the analysis of microarray data," *J. Comput. Biol.*, vol. 11, pp. 1090–1109, 2004.
[21] C. H. Zheng, D. S. Huang, K. Li, G. Irwin, and Z. L. Sun, "MISEP method for post-nonlinear blind source separation," *Neural Comput.*, vol. 19, pp. 2557–2578, 2007.
[22] D. S. Huang and C. H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," *Bioinformatics*, vol. 22, pp. 1855–1862, 2006.
[23] S. I. Lee and S. Batzoglou, "Application of independent component analysis to microarrays," *Genome Biol.*, vol. 4, p. R76, 2003.
[24] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.
[25] J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM J. Matrix Anal. Appl.*, vol. 17, pp. 161–164, 1996.
[26] S. A. Saidi, C. M. Holland, D. P. Kreil, D. J. C. Mackay, D. S. Charnockjones, C. G. Print, and S. K. Smith, "Independent component analysis of microarray data in the study of endometrial cancer," *Oncogene*, vol. 23, pp. 6677–6683, 2004.
[27] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.
[28] Z. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized parts-based representations," in *Proc.IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, Hawaii, 2001, vol. 1, pp. 207–212.
[29] P. O. Hoyer, "Non-negative sparse coding neural networks for signal processing," in *Proc. IEEE Workshop Neural Netw. Signal Process.*, Martigny, Switzerland, 2002, pp. 557–565.
[30] F. Shahnaz, M. Berry, P. Pauca, and R. Plemmons, "Document clustering using nonnegative matrix factorization," *J. Inf. Process. Manag.*, vol. 42, pp. 373–386, 2006.
[31] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. (NIPS 2000)*. vol. 13, Cambridge, MA: MIT Press, 2001, pp. 556–562.
[32] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: A resampling-base method for class discovery and visualization of gene expression microarray data," *Mach. Learn.*, vol. 52, pp. 91–118, 2003.
[33] R. R. Sokal and F. J. Rohlf, "The comparison of dendrograms by objective methods," *Taxon*, vol. 11, pp. 33–40, 1962.
[34] Cophenetic correlation. [Online]. Available: http://en.wikipedia.org/wiki/Cophenetic_correlation
[35] W. Xu, X. Liu, and Y. Gong, "Document-clustering based on non-negative matrix factorization," in *Proc. SIGIR 2003*, Toronto, CA, Jul. 28–Aug. 1, pp. 267–273.
[36] H.-Q. Wang, H.-S Wong, D. S. Huang, and J. Shu, "Extracting gene regulation information for cancer classification," *Pattern Recognit.*, vol. 40, pp. 3379–3392, 2007.
[37] S. L. Wang, H. W. Chen, F. Li, and D. Zhang, "Gene selection with rough sets for the molecular diagnosing of tumor based on support vector machines," in *Proc. Int. Comput. Symp. 2006*, Taiwan, pp. 1368–1373.
[38] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.
[39] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in *A Practical Approach to Microarray Data Analysis*. Norwell, MA: Kluwer, 2003, pp. 91–109.



Chun-Hou Zheng was born in Shandong, China, in 1973. He received the B.Sc. degree in physics education and the M.Sc. degree in control theory and control engineering from Qufu Normal University, Rizhao, China, in 1995 and 2001, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China, in 2006.

From February 2007 to March 2009, he was a Postdoctoral Fellow in the Intelligent Computing Laboratory, Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei. He is currently an Associate Professor in the College of Information and Communication Technology, Qufu Normal University. His current research interests include intelligent computing and bioinformatics.

De-Shuang Huang (SM'98) received the B.Sc. degree from the Institute of Electronic Engineering, Hefei, China, in 1986, the M.Sc. degree from the National Defense University of Science and Technology, Changsha, China, in 1989, and the Ph.D. degrees from Xidian University, Xian, China, in 1993, all in electronic engineering.

From 1993 to 1997, he was a Postdoctoral Fellow at Beijing Institute of Technology. From 1993 to 1997, he was also a Postdoctoral Fellow at the National Key Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China, where he is currently with the Intelligent Computing Laboratory, Institute of Intelligent Machines and was a recipient of the "Hundred Talents Program of CAS" in September 2000. From September 2000 to March 2001, he was a Research Associate at Hong Kong Polytechnic University, where he was also a Research Fellow from October to December 2003. From April 2002 to June 2003, he was a Research Fellow at the City University of Hong Kong. From August to September 2003, he was a Visiting Professor at George Washington University, Washington, DC. From July to December 2004, he was a University Fellow at Hong Kong Baptist University. From March 2005 to March 2006, he was a Research Fellow at Chinese University of Hong Kong. From March to July 2006, he was a Visiting Professor at Queen's University of Belfast, U.K. From October to November 2007, he was a Visiting Professor at Inha University, Korea. He has authored or coauthored more than 200 papers. He has authored or coauthored two books: *Systematic Theory of Neural Networks for Pattern Recognition* (1996) and *Intelligent Signal Processing Technique for High Resolution Radars* (2001).

Dr. Huang received the Second-Class Prize of the 8th Excellent High Technology Books of China for his book *Systematic Theory of Neural Networks for Pattern Recognition*.

Lei Zhang (M'04) received the B.S. degree from Shenyang Institute of Aeronautical Engineering, Shenyang, China, in 1995, the M.Sc. and Ph.D. degrees in electrical and engineering from Northwestern Polytechnical University, Xi'an, China, in 1998 and 2001, respectively.

From 2001 to 2002, he was a Research Associate in the Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong, where has been an Assistant Professor since January 2006. From January 2003 to January 2006, he was a Postdoctoral Fellow in the Department of Electrical and Computer Engineering, McMaster University, Canada. His current research interests include image and video processing, biometrics, bioinformatics, pattern recognition, multisensor data fusion and optimal estimation theory, etc.

Xiang-Zhen Kong was born in 1979. She received the M.Sc. degree from the Institute of Automation, Qufu Normal University, Qufu, China, in 2008.

She is currently a Lecturer in the College of Information and Communication Technology, Qufu Normal University. Her current research interests include artificial neural networks and intelligent information processing.