Metasample Based Sparse Representation for Tumor Classification

Chun-Hou Zheng, Lei Zhang, To-Yee Ng, and Chi Keung Shiu

Abstract: A reliable and accurate identification of the type of tumors is crucial to the proper treatment of cancers. In recent years, it has been shown that sparse representation (SR) by l_1-norm minimization is robust to noise, outliers and even incomplete measurements, and SR has been successfully used for classification. This paper presents a new SR based method for tumor classification using gene expression data. A set of metasamples are extracted from the training samples, and then an input testing sample is represented as the linear combination of these metasamples by l_1-regularized least square method. Classification is achieved by using a discriminating function defined on the representation coefficients. Since l_1-norm minimization leads to a sparse solution, the proposed method is called metasample based SR classification (MSRC). Extensive experiments on publicly available gene expression datasets show that MSRC is efficient for tumor classification, achieving higher accuracy than many existing representative schemes.

Index Terms: Tumors Classification, Sparse Representation, Metasample, Gene Expression Data.

C.-H. Zheng is with the College of Information and Communication Technology, Qufu Normal University, Rizhao 276826, China, and with the Biometric Research Center, Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: zhengch99@126.com).

L. Zhang is with the Biometric Research Center, Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China (Corresponding author; phone: 852-27667355; e-mail: cslzhang@comp.polyu.edu.hk).

T.-Y. Ng is with the Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: cstyng@comp.polyu.edu.hk;).

C.-K. Shiu is with the Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: csckshiu@comp.polyu.edu.hk).

1. Introduction

A tumor is a neoplasm or a solid lesion formed by an abnormal growth of cells. A reliable and accurate identification of the type of tumors is crucial to the effective treatment of cancers. Tumor is different from cancer. A tumor can be benign, pre-malignant or malignant, while only the malignant tumor can be called cancer. The management of cancer has been attracting tremendous attention because cancer is a potentially life-threatening disease caused by the unchecked proliferation of cells.

DNA microarray is a biotechnology that simultaneously monitors the expression of tens of thousands genes in cells [1]. One important and emerging application of microarray gene expression profiling is tumor classification [2,3]. Many classification methods originated from statistical learning theory have been adapted for molecular data classification or clustering [3-10,25,26,39]. Golub *et al.* [3] successfully classified acute myeloid leukemia from acute lymphocytic leukemia using gene expression data. Huang *et al.* [9] proposed to used independent component analysis (ICA) based penalized discriminant method for tumor classification. Brunet used nonnegative matrix factorization (NMF) to cluster tumor samples. A major methodological concern of such methods is the problem of over-fitting, i.e., method become over-optimized to perform well on the training set, but does not generalize well to new data from the same class of cancer [11].

Sparse representation (SR) is a new and powerful data processing method, which is inspired by the recent progress of l_1-norm minimization based methods such as basis pursuing [12], compressive sensing for signal reconstruction [13-15], and least absolute shrinkage and selection operator (LASSO) algorithm for feature selection [16]. By using the SR technique to represent the input testing face image as a sparse linear combination of the training samples, an SR based classification (SRC) method was proposed in [18] for face recognition. Ideally, in SRC it is expected that a testing sample

can be well represented by using only the training samples from the same class. In such case, the SR coefficient vector will have only a few significant coefficients. In order to find the SR coefficient vector, l_1-regularized least square [17] is used. Unlike conventional supervised learning methods, where a training procedure is used to create a classification model for testing, the SRC does not contain separate training and testing stages so that the over-fitting problem is much lessened. The SRC method has been successfully used in face recognition [18] and tumor classification [19]. In these methods, a testing sample is represented as the linear combination of the original training samples, and the representation error over each class is used as an indicator to classify the testing sample. However, the original training samples may be not as efficient as the eigenfaces [20] or metasamples [5], which contain the intrinsic structural information of the data, to represent the input testing samples.

A *metasample* is a linear combination of the gene expression profiles of samples, which can capture the alternative structures inherent to the data. The samples are analyzed by summarizing their gene expression patterns in terms of expression patterns over the metasamples. In [5, 21], the metasample expression patterns discovered by NMF provide a robust clustering of samples[*]. In [22], the authors proposed a similar method called *Eigenarrays*, which are extracted by using singular value decomposition (SVD) or principal component analysis (PCA) from the gene expression data. For the convenience of expression, we use the word *metasample* throughout the paper. The metasamples can be computed by using SVD, PCA, NMF or other linear or nonlinear models such as ICA and nonlinear ICA [23,24].

In this study we propose to represent each testing sample as a linear combination of a set of

[*] In [5], the authors used the term “metagene” to represent metasample.

metasamples extracted from all the training samples. Classification is achieved by using a discriminating function of the representation coefficients on the metasamples obtained by l_1-regularized least square. Since l_1-norm minimization could lead to sparse solution, our approach is then named as metasample based sparse representation classification (MSRC).

The rest of the paper is organized as follows. Section 2 describes the methods proposed in this study. The SR of tumor samples and the metasample model of gene expression data are first presented, and the algorithm of MSRC is consequently given. Section 3 presents the numerical experiments. Section 4 concludes the paper and outlines directions of future work.

The abbreviations used in this paper are summarized as follows.

Table 1. Abbreviations

SR	sparse representation
SRC	SR based classification
MSRC	metasample based SRC
NMF	nonnegative matrix factorization
SNMF	sparse NMF
SVD	singular value decomposition
PCA	principal component analysis
ICA	independent component analysis
SVM	support vector machines
KNN	K-nearest neighbors
GEMS	gene expression model selector
BW	between-categories to within-category
LASSO	least absolute shrinkage and selection operator

2. Methods

2.1 Sparse Representation of Testing Tumor Samples

The basic problem of supervised tumor classification is how to use the labeled training samples from the k object classes to correctly identify the class to which a new testing sample belongs. Consider a training gene expression dataset represented by an $m \times n$ matrix A with m genes and n samples.

Since the microarray data typically contain thousands of genes on each chip, and the number of collected tumor samples is much smaller than that of genes, we have $m \gg n$. The element a_{ql} in A is the expression level of the q-th gene in the l-th assay ($1 \le q \le m$, $1 \le l \le n$). The n-dimensional vector r_q, i.e., the q-th row of A, denotes the expression profile of the q-th gene. Alternatively, the m-dimensional vector c_l, i.e., the l-th column of A, is the snapshot of the l-th assay (cell sample). Here we suppose that the n samples belong to k object classes ($k \le n$).

We arrange the n_i samples of the i-th class ($1 \le i \le k$) as a matrix $A_i = [\mathbf{c}_{i,1}, \mathbf{c}_{i,2}, \cdots \mathbf{c}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$ with each sample being a column. Given that the training samples of the i-th class are sufficient, any new (testing) sample $y \in \mathbb{R}^m$ in the same class will approximately lie in the linear span of the training samples associated with class i [18]:

$$y = \alpha_{i,1}\mathbf{c}_{i,1} + \alpha_{i,2}\mathbf{c}_{i,2} + \cdots + \alpha_{i,n_i}\mathbf{c}_{i,n_i} \tag{1}$$

for some scalars $\alpha_{i,j} \in \mathbb{R}$, $j = 1, 2, \cdots, n_i$.

For tumor classification, the membership i of the new testing sample y is unknown. We arrange the training data samples of each class in matrix A. Suppose that the samples with the same class are conjoint, i.e., $A = [A_1, A_2, \cdots A_k]$, then the linear representation of y can be rewritten in terms of all the training samples as

$$y = Ax_0 \tag{2}$$

where, ideally,

$$x_0 = [0, \cdots 0, \alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,n_i}, 0, \cdots, 0]^T \in \mathbb{R}^n \tag{3}$$

is a coefficient vector whose entries are zero except for those associated with the i-th class. In other words, the nonzero entries in the estimate x_0 will be associated with the columns of A from a single object class i so that we can assign the testing sample y to that class.

Now, the key problem to be shoved is how to calculate x_0. From Eq. (3) we can see that the representation of y is naturally sparse if the number of object classes k is large. The problem can

be converted into how to find a column vector x such that $y = Ax$ and $\|x\|_0$ is minimized, where $\|x\|_0$ is the l_0-norm of x and it is equivalent to the number of nonzero elements in vector x, i.e., the so-called sparse representation (SR). It can be expressed as the following optimization problem:

$$\hat{x}_0 = \arg\min \|x\|_0 \quad \text{subject to} \quad Ax = y \tag{4}$$

Finding the solution to the above SR problem is NP-hard due to its nature of combinational optimization [27]. Fortunately, recent development in the theory of SR and compressive sensing [13-16] reveals that if the solution being sought is sparse enough, the solution to the l_0- minimization problem in Eq. (4) is equivalent to the solution to the following l_1-minimization problem:

$$\hat{x}_1 = \arg\min \|x\|_1 \quad \text{subject to} \quad Ax = y \tag{5}$$

This problem can be solved in polynomial time by standard linear programming methods [12]. For A, whose size $m >> n$, Eq. (5) does not have exact solutions. To solve this problem, we consider a generalized version of Eq. (5), which allows for certain degree of noise, i.e., find a vector x such that the following objective function is minimized:

$$J(x, \lambda) = \min_x \left\{ \|Ax - y\|_2 + \lambda \|x\|_1 \right\} \tag{6}$$

Using this, Eq. (5) is reduced to solving an l_1-regularized least square problem. The positive parameter λ in Eq. (6) is a scalar regularization that balances the reconstruction error and sparsity. This optimization problem can also be solved by standard linear programming methods [12]. In this study, we use the truncated Newton interior-point method [17] to solve this problem.

2.2 Metasample of Gene Expression Data

Generally speaking, metasample of gene expression data is defined as a linear combination of several samples, which may capture alternative structures inherent to the data and provide biological insight. Another viewpoint of metasample is that we can approximate the gene expression pattern as linear

combinations of these metasamples. Mathematically, the gene expression dataset matrix A can be factorized into two matrices

$$A \sim WH, \tag{7}$$

where matrix W is of size $m \times p$, with each of the p columns defining a metasample. Matrix H is of size $p \times n$, with each of the n columns representing the metasample expression pattern of the corresponding sample (refer to Fig.1).

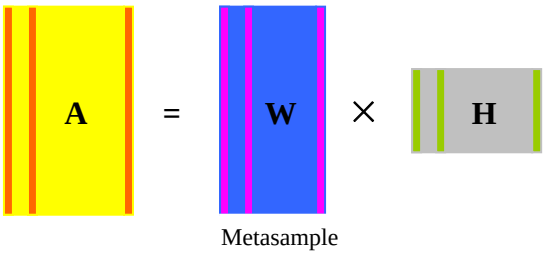

Fig. 1 The metasample model of gene expression data. Each sample (column in A) in the data matrix is considered to be a linear combination of underlying basis snapshots (metasamples) in the matrix W (columns in W). Each column in H represents the metasample expression pattern of the corresponding sample.

Many works have been published on extracting the metasamples [5, 22, 23]. Alter *et al.* [22] used SVD to transform the gene expression data from the "genes × samples" space to diagonalized "eigengenes × eigenarrays (i.e., metasamples)" space, where the eigengenes (or eigenarrays) are unique orthonormal superposition of the genes (or samples). They found that sorting the data according to the eigengenes and eigenarrays gives a global picture of the dynamics of gene expression, in which individual genes and arrays appear to be classified into groups of similar regulation and function, or similar cellular state and biological phenotype, respectively. After normalization and sorting, the significant eigengenes and eigenarrays can be associated with observed genome-wide effects of regulators, or with measured samples, in which these regulators are overactive or underactive, respectively.

In our previous works [9, 25], ICA was used to model gene expression data as shown in Fig 1. In this model, the samples are considered to be a linear mixture of statistically independent basis snapshots (i.e., metasamples). On the other hand, Brunet et al. [5] used NMF to describe the gene expression data in terms of a small number of metasamples. Then the samples are analyzed by summarizing their gene expression patterns in terms of expression patterns of the metasamples. It was shown in [5] that the metasample expression patterns provide a robust clustering of samples.

Form the above analysis we see that the metasamples can be extracted using several methods, such as SVD, ICA and NMF, etc. Based on the method used, different names were proposed, including eigengene, eigenarray, independent basis snapshot and metasample. We choose the name "metasample". In the following section, we use SVD and sparse NMF (SNMF) to extract the metasamples for classification because of their efficiency.

2.3 Metasample based Sparse Representation Classification

Since the metasamples contain the inherent structural information of training samples, in this study we propose to use them to design the classifier instead of the original samples in training dataset. We extract the metasamples from the samples in each class respectively, i.e., we factorize each sub-dataset matrix A_i into two matrices:

$$A_i \sim W_i H_i \tag{8}$$

where W_i is an $m \times p_i$ matrix and H_i is a $p_i \times n_i$ matrix. p_i is the number of metasamples of the *i*-th class. In practice, the value of p_i can be determined experimentally. In this study, we use SVD and SNMF to solve Eq.(8) for W_i.

After computing the metasamples W_i of each class, we use W to represent the metasamples from all the *k* classes:

$$W \doteq [W_1, W_2, \cdots, W_k] \tag{9}$$

Given a new test sample y, we can compute its SR by minimizing

$$J(x, \lambda) = \min_x \left\{ \|Wx - y\|_2 + \lambda \|x\|_1 \right\}, \tag{10}$$

where the positive scalar regularization parameter λ is determined experimentally.

Ideally, the nonzero entries in the representation vector *x* will all be associated with the columns of *W* from a single class i, allowing us to assign the testing sample *y* to the class *i*. However, noise and modeling error will inevitably lead to small nonzero entries associated with multiple object classes [18]. To solve this problem and for a more robust classification, we classify *y* based on how well *y* can be reconstructed by using the coefficients from each class as in [18].

For each class i, let $\delta_i : \mathbb{R}^n \to \mathbb{R}^n$ be the characteristic function which selects the coefficients associated with the *i*-th class. For $x \in \mathbb{R}^n$, $\delta_i(x) \in \mathbb{R}^n$ is a vector whose nonzero entries are the ones from class *i* in x. Using only the coefficients from the *i*-th class, one can reconstruct the given test sample y as $\hat{y}_i = W\delta_i(x)$. We then classify y based on these approximations by assigning it to the class that minimizes the residual between y and $\hat{y}$:

$$\min_i r_i(y) \doteq \|y - W\delta_i(x)\|_2 \tag{11}$$

The classification algorithm can be summarized as following:

Input: matrix of training samples $A = [A_1, A_2, \cdots, A_k] \in \mathbb{R}^{m \times n}$ for k classes; testing sample $y \in \mathbb{R}^m$.

Step1: Normalize the columns of *A* to have unit l_2-norm.

Step2: Extract the metasamples of every class using SVD or NMF.

Step3: Solve the optimization problem defined in Eq. (10).

Step4: Compute the residuals $r_i(y) = \|y - W\delta_i(x)\|_2$.

Output: identity(y) = $\arg\min_i r_i(y)$.

Since our algorithm is based on metasamples, we name it the metasample based sparse representation classification (MSRC). The optimization problem in Eq. (10) is solved using the truncated Newton interior-point method, which is done by l1_ls MATLAB package available online (http://www.stanford.edu/~boyd/l1_ls). The Matlab codes of the proposed MSRC algorithm can be downloaded at http://www4.comp.polyu.edu.hk/~cslzhang/code/MSRC_IEEE.rar.

MSRC can be seen as the combination of SRC [18] and metasample based clustering [5]. In SRC, the testing sample is represented as a linear combination of the original training samples. In metasample based clustering, each sample is represented as a linear combination of metasamples, which are extracted from the training samples. The common point of the two methods is that they are both using the coefficient vector for classification or clustering. The difference between them is that in the proposed MSRC, the testing sample is represented as a linear combination of metasamples extracted in a supervised manner from each class separately.

2.4 Evaluation of Performance

The proposed method is evaluated in comparison with some representative methods, including the SRC [18, 19], LASSO [40] and the widely used support vector machines (SVM) [28]. SVM has been successfully used for gene profile classification [28]. Considering the characteristics of 'high dimensionality and small sample size' of gene expression data, SVM may be the best classifier for classifying the original data [29, 30]. Statnikov et al. [29] and Pochet et al. [30] compared various methods for tumor classification and concluded that SVM is among the most efficient ones, outperforming K-nearest neighbors (KNN) and neural network. Based on this conclusion, KNN and

neural network will not be used in the comparison, though they are also useful for tumor classification. In addition, it has been reported that SRC has similar performance to SVM in classification [18].

The experiments of two-class classification are given in subsection 3.1. We use classification accuracy, sensitivity and specificity, as the performance metrics. They are obtained by stratified 10-fold cross-validation. The results of SVM are obtained by the Gene Expression Model Selector (GEMS), a set of software with graphic user interface for gene expression data classification. GEMS is publically available at http://www.gems-system.org/ and it was also used in [29] for the comprehensive study of many classifiers on gene expression cancer diagnosis. In this study, we use one-versus-rest SVM (OVRSVM) with Polynomial kernels to do the experiment because Statinkov et al. have shown that OVRSVM may be the best one for tumor classification [29].

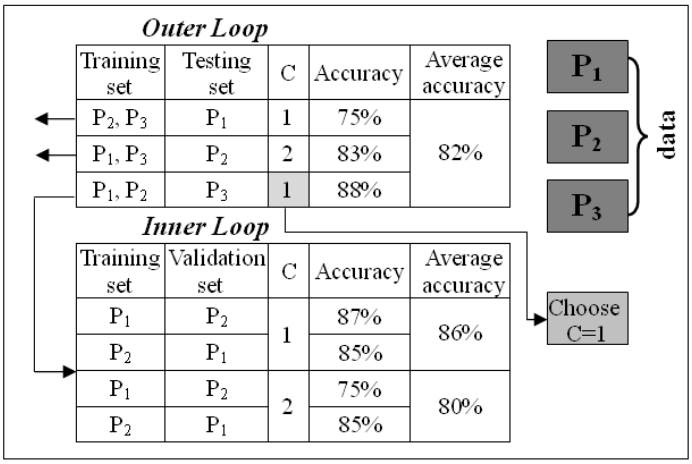

Fig.2. An example of parameter selection. Data are split into mutually exclusive sets P_1, P_2 and P_3. The performance is estimated in the outer loop by training on all splits but one, and using the remaining one for testing. The average performance over testing sets is reported. The inner loop is used to determine the optimal value of parameter C (in a cross-validated fashion) for training in the outer loop.

Nested stratified 10-fold cross validation method [29] is used to select the parameters of the classifiers, which is based on two loops. The inner loop is used to determine the best parameter of the classifier. The outer loop is used for estimating the performance of the classifier. Fig. 2 shows a

simplified pictorial example of a nested stratified 3-fold cross validation applied to optimize the parameter *C* (which takes values '1' and '2') of a classifier. Note that in reality we will optimize a set of combined parameters.

The multi-class classification experiments are given in subsection 3.2. Because the LASSO based method is designed for two-class classification problem [40], we do not use it for multi-class tumor classification. The numbers of training samples in different classes are unbalanced, e.g., in the lung cancer dataset, there are 139 adenocarcinoma samples but only 6 small-cell lung carcinoma samples. Thus in the experiments we can set different numbers of metasamples according to the numbers of samples in different classes, i.e., more metasamples will be extracted if the number of training samples of that class is bigger. However, the classification results may be biased towards classes with greater numbers of metasamples. On the other hand, if the number of metasamples of each class is set the same, it must be equal to or less than the smallest n_i, and thus we will lose the information contained in larger classes. With the above considerations and according to the results of nested stratified 10-fold cross validation, we choose $p_i = 8$ if $n_i > 8$. Otherwise, we let $p_i = n_i$.

To study whether dimensionality reduction can improve the classification performance, we also applied SRC, LASSO, SVM and MSRC to subsets of selected top-ranked genes. It should be noted that since the number of genes are very large, an optimal selection of the genes is very computationally expensive. Considering that our aim is to validate whether gene selection can benefit MSRC and the fact that cancer is caused by the mutation of many genes, we use a simple method, i.e., the ratio of genes between-categories to within-category sums of squares (BW) method, to rank the genes, and select a comparatively large number of genes for the experiment.

3. Results

3.1 Two-Class Classification

Table 2. Summary of the datasets for the four binary tumor classification problems.

Datasets	Samples		Genes
	Class 1	Class 2	
Acute leukemia data	19	19	5000
Colon data	40	22	2000
Prostate cancer data	77	59	12600
DLBCL	58	19	5469

Table 3. The classification accuracies by different methods.

Dataset	SVM	LASSO	SRC	MSRC-SNMF	MSRC-SVD
Acute leukemia	**97.37**	86.84	94.74	**97.37**	**97.37**
Colon	85.48	85.48	85.48	**90.32**	**90.32**
Prostate	91.18	91.91	94.85	91.91	**95.59**
DLBCL	96.10	96.10	97.40	97.40	**98.70**

Table 4. The classification sensitivity by different methods.

Dataset	SVM	LASSO	SRC	MSRC-SNMF	MSRC-SVD
Acute leukemia	**94.74**	89.47	89.47	**94.74**	**94.74**
Colon	92.50	90.00	**92.70**	92.50	92.50
Prostate	93.51	90.91	93.51	90.91	**94.81**
DLBCL	**98.28**	**98.28**	**98.28**	**98.28**	**98.28**

Table 5. The classification specificity by different methods.

Dataset	SVM	LASSO	SRC	MSRC-SNMF	MSRC-SVD
Acute leukemia	**100**	84.21	**100**	**100**	**100**
Colon	72.73	77.27	72.73	**86.36**	**86.36**
Prostate	88.14	93.22	96.61	93.22	**96.61**
DLBCL	89.47	89.47	94.74	94.74	**100**

Four publicly available microarray datasets are used to study the tumor classification problem: Acute leukemia dataset [3], Colon cancer dataset [4], Prostate cancer dataset [37] and Diffuse large B-cell lymphomas (DLBCL) dataset [36].

The leukemia dataset is consisted of 38 bone marrow samples, which were obtained from adult

acute leukemia patients at the time of diagnosis and before chemotherapy. RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays to measure the gene expression values. For the colon dataset, gene expression in 40 tumor and 22 normal colon tissue samples was analyzed with an Affymetrix oligonucleotide array complementary to more than 6500 human genes. The data set contains the expression of 2000 genes with the highest minimal intensity across 62 tissues. For the prostate dataset, the gene expression profiles were derived from prostate tumors and non-tumor prostate samples from patients undergoing surgery. Oligonucleotide microarrays containing probes for approximately 12600 genes and ESTs were used for obtaining the gene expression data. For the DLBCL dataset, RNA was hybridized to high-density oligonucleotide microarrays to measure the gene expression values. An overview of the characteristics of the four datasets is given in Table 2.

The classification results (including accuracy, sensitivity and specificity) by using SVM, LASSO, SRC and the proposed MSRC are listed in Tables 3-5. SVD and SNMF are used to extract the metasamples of gene expression data, respectively. When using the two methods to extract the metasamples, we need to determine the number of metasamples of each class, i.e., the value of p_i in Eq. (8). Since there are only 2 classes and the difference of the number of samples in each class is not big, we let $p_1 = p_2 = p$. The value of p can be determined using the nested stratified 10-fold cross validation. Another parameter λ in Eq.(10) can also be determined using this method. Note that, the results of SVM and SRC in our experiments are slightly different from those reported in [19, 29]. This is probably because the distribution file of cross validation in our experiments is different from those in [19, 29].

To better illustrate the experimental results, we show the accuracies of our methods MSRC-SVD

and MSRC-SNMF in Figures 3-6 when different numbers of metasamples are used. The *x*-axis represents the *p*-dimension, i.e., the number of metasamples extracted from the original data; *y*-axis represents the accuracy of classification.

From Tables 3-5 and Figures 3-6 it can be seen that for the four datasets, MSRC-SVD achieves better classification results than SVM and SRC, especially on the colon and DLBCL datasets. The proposed MSRC-SVD is very competitive for binary tumor classification. From Figures 3-6 we can also find that, compared with MSRC-SNMF, MSRC-SVD could achieve higher accuracy. This may be in conflict with the intuitive sense that SNMF could model the gene expression data in a more biological way [8]. From another viewpoint, however, SVD is the optimal model for reconstruction under l_2-norm. Certainly, more experiments could be performed to further validate this in the future.

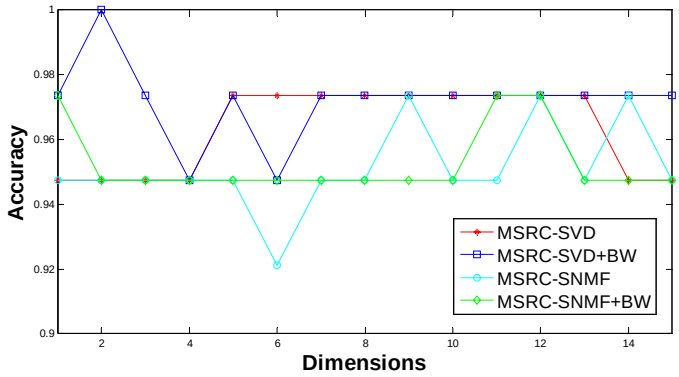

Fig.3. The classification accuracy on the Acute leukemia dataset.

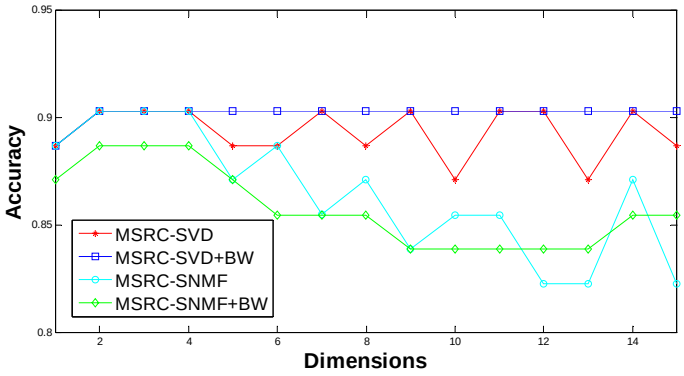

Fig.4. The classification accuracy on the colon dataset.

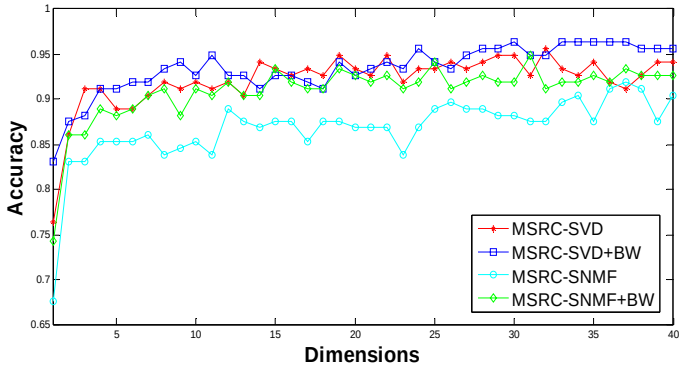

Fig.5. The classification accuracy on the Prostate cancer dataset.

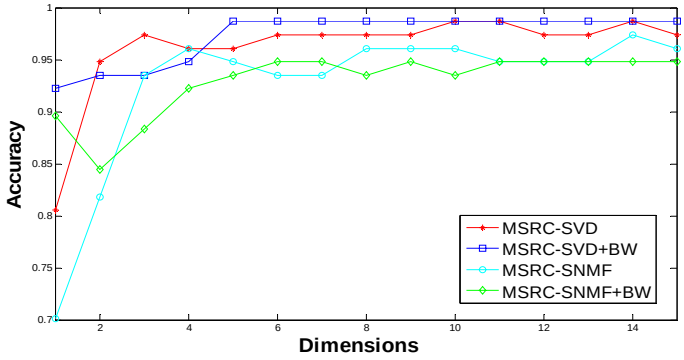

Fig.6. The classification accuracy on the DLBCL dataset.

Table 6. The classification accuracies by different methods (with gene selection).

Dataset	SVM	LASSO	SRC	MSRC-SNMF	MSRC-SVD
Acute leukemia (2000*)	97.37	89.47	97.37	97.37	**100**
Colon (1000)	87.10	87.10	87.10	88.71	**90.32**
Prostate (1500)	94.85	91.18	95.59	94.85	**96.32**
DLBCL (800)	97.40	93.51	97.40	94.81	**98.70**

* The number of selected top-ranked genes.

Table 7. The classification sensitivity by different methods (with gene selection).

Dataset	SVM	LASSO	SRC	MSRC-SNMF	MSRC-SVD
Acute leukemia	94.74	94.74	94.74	94.74	**100**
Colon	**92.50**	**92.50**	90.00	**92.50**	**92.50**
Prostate	92.21	90.91	94.81	92.21	**94.81**
DLBCL	**98.28**	96.55	**98.28**	93.10	**98.28**

Table 8. The classification specificity by different methods (with gene selection).

Dataset	SVM	LASSO	SRC	MSRC-SNMF	MSRC-SVD
Acute leukemia	**100**	84.21	**100**	**100**	**100**
Colon	77.27	77.27	81.82	81.82	**86.36**
Prostate	**98.31**	91.53	96.61	**98.31**	**98.31**
DLBCL	94.74	84.21	94.74	**100**	**100**

The experimental results with gene selection are shown in Tables 6-8 and Figures 3-6 (MSRC-SVD+BW and MSRC-SNMF+BW). We can see that, except for MSRC-SNMF and LASSO, gene selection can improve the performance of all the other classification methods. For MSRC-SVD, gene selection can improve both the accuracy and the stability of the classification. For MSRC-SNMF and LASSO, no clear regularity could be found.

3.2 Multi-class Classification

We use five multi-class datasets to further investigate the performance of the proposed method. 1) The Lung cancer dataset [34], which contains 4 lung cancer types and normal tissues (i.e., 5 classes in total). This dataset includes 203 samples with 12600 genes. 2) The Leukemia dataset [33], which has three kinds of samples: acute myelogenous leukemia, acute lymphoblastic leukemia and mixed-lineage leukemia. This dataset includes 72 samples with 11225 genes. 3) The Small round blue cell tumors of childhood (SRBCT) [35], which is composed of 4 types of tumors. This dataset includes 83 samples with 2308 genes. 4) A dataset composed of 11 various human tumor types (11_Tumors [32]): ovary, bladder/ureter, breast, colorectal, gastro-esophagus, kidney, liver, prostate, pancreas, adeno lung, and squamous lung. This dataset includes 174 samples with 12533 genes. 5) A dataset composed of 9 various human tumor types (9_Tumors [31]): non-small cell lung, colon, breast, Ovarian, Leukemia, Renal, Melanoma, Prostate, central nervous system. This dataset includes 60

samples with 5726 genes.

All the five datasets were produced by oligonucleotide microarrays. Except SRBCT, for the other four datasets RNA was hybridized to high-density oligonucleotide Affymetrix arrays, and expression values were computed using the analysis tool: Affymetrix GENECHIP [29]. The SRBCT dataset was produced by using two-color cDNA platform with consecutive image analysis performed by the DeArray Software and filtering for a minimal level of expression [35].

The experimental results are listed in Table 9. From the five experiments we can see that, for multi-class classification the proposed MSRC does not have clear advantages over SVM and SRC. The reason is that in these datasets, the training samples of some classes are very few so that the extracted metasamples cannot represent the intrinsic information of these classes. For example, the number of samples of each class in the 9_tumors dataset is listed in Table 10. We see that the samples of each class are very few, which makes the classification result of MSRC relatively poor on this dataset.

Table 11 shows the experimental results with gene selection. Compared with Table 9, we can see that except for the 9_tumors dataset, little improvement is achieved by gene selection for the other three datasets. Since the accuracy on SRBCT dataset is 100% for all the classifiers, we did not apply gene selection on this dataset.

Table 9. The multi-class classification accuracies by different methods.

Dataset	SVM	SRC	MSRC-NMF	MSRC-SVD
Lung cancer	**96.05**	95.07	94.09	95.07
Leukemia	96.90	95.83	95.83	**97.22**
SRBCT	100	100	100	100
11_tumors	94.68	94.83	95.40	**95.98**
9_tumors	65.10	**66.67**	60.60	63.33

Table 10. The sample numbers of every class in 9_tumor dataset.

Tissues	NSCL	Colon	Breast
Numbers	9	7	8
Tissues	Ovarian	Leukemia	Renal
Numbers	6	6	8
Tissues	Melanoma	Prostate	CNS
Numbers	8	2	6

Table 11. The multi-class classification accuracies (with gene selection).

Dataset	SVM	SRC	MSRC-NMF	MSRC-SVD
Lung cancer (2000)	**96.62**	95.07	94.09	96.06
Leukemia (3000)	96.90	95.83	95.83	**97.22**
11_tumors (1000)	96.07	95.40	93.10	**96.55**
9_tumors (2000)	**85.84**	71.67	66.67	73.33

3.3 The Required Number of Samples for Metasample Training

From the aforementioned experimental results we see that our method could efficiently classify tumor data, especially when there are enough training samples of each class. On the other hand, if the training samples are very few, our method may not be better than SVM and SRC.

To find out how many training samples are required by MSRC to perform better than SRC, we randomly select different numbers of samples from the four two-class datasets and the Leukemia dataset [33] for testing. Each class in these datasets has more than 18 samples. We randomly chose 6, 10, 14 and 18 samples, respectively, from each class of the dataset for classification. Each experiment was repeated 10 times. The mean classification accuracies of 10-fold cross-validation are listed in Table 12. From this table we can see that MSRC will not have clear advantages over SRC when the number of training samples is less than 10. If there are 10 or more than 10 training samples, MSRC will be a good choice for tumor classification. This is consistent to the results in subsection 3.2.

Table 12. The mean classification accuracies on the randomly selected subsets.

Number of samples for each class	Dataset	Accuracy(%)	
		SRC	MSRC
6 samples	Acute leukemia	89.16	89.16
	Colon	90.83	90.83
	Prostate	84.99	84.99
	DLBCL	90.00	90.00
	Leukemia	91.66	89.44
10 samples	Acute leukemia	93.00	94.50
	Colon	84.50	88.00
	Prostate	87.50	89.50
	DLBCL	95.00	95.00
	Leukemia	93.33	93.33
14 samples	Acute leukemia	96.07	96.43
	Colon	84.92	86.42
	Prostate	82.14	85.35
	DLBCL	95.71	97.14
	Leukemia	95.95	96.66
18 samples	Acute leukemia	93.60	94.71
	Colon	82.22	86.66
	Prostate	87.22	88.88
	DLBCL	93.88	95.55
	Leukemia	95.74	96.66

4. Conclusion and Discussion

Cancer diagnosis is one of the most important emerging clinical applications of gene expression data. In this study, we proposed a new sparse representation (SR) based approach for cancer diagnosis, which expresses each testing sample as a linear combination of a set of metasamples extracted using SVD or NMF from the training samples. Classification is achieved by a discriminating function of the SR coefficients, which are obtained by l_1-regularized least square optimization. The proposed metasample based SR classification (MSRC) was compared with the standard SR classification (SRC) and the benchmark SVM methods on 9 typical datasets. The results validated that MSRC is effective and efficient in tumor classification. Since it is not comprised of training and testing process, as other

classifiers, e.g., SVM, our method (including SRC) has no over-fitting problem.

The experimental results also show that, compared with SRC, MSRC is a better choice if there are 10 or more than 10 training samples. The reason may be that when there are 10 or more than 10 training samples, metasamples can capture the intrinsic structural information of the data of each class, and thus MSRC shows superior classification performance to SRC. On the other hand, if the number of training samples is less than 10, the trained metasamples may not be able to capture sufficient intrinsic structural information of each class, and the performance of MSRC is similar to or slightly worse than SRC. This is one weakness of the proposed method, i.e. the training samples for metasample training cannot be too limited. What should be noted is that, our method is based on the hypothesis that the testing sample can be well represented as a linear combination of the training samples from the same class. If the hypothesis is invalid, our method will not work. Fortunately, the success of sparse representation in face recognition [18] and gene expression data classification [19] has demonstrated that this hypothesis holds well. Certainly, more experiments on more databases need to be performed in the future to further validate the proposed method.

Since gene selection can enhance the accuracy of the classification [29], we also used gene selection to preprocess the gene expression data. As expected in theory, the experimental results showed that gene selection can enhance the performance of SVM, SRC, as well as MSRC-SVD. For MSRC-NMF, no clear rule can be found. In the future, we will investigate whether other metasample extraction methods, such as sparse PCA, KPCA, etc., will achieve better results. In addition, we will also investigate how to solve the metasample extraction and the testing sample classification as one unified optimization problem, which may improve the accuracy of classification.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under grant no. 30700161 and the Foundation for Young Scientist of Shandong Province, China under grant no. 2008BS01010.

References

[1] P.O. Brown, and D. Botstein, “Exploring the new world of the genome with DNA microarray,” The Chipping Forest, vol.21, pp.33–37, 1999.

[2] E.E. Ntzani, and J.P. Ioannidis, “Predictive ability of DNA microarrays for cancer outcomes and correlates: and empirical assessment,” Lancet, vol.362, pp.1439–1444, 2003.

[3] T.R. Golub, D.K. Slonim, P.Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov *et al.*, “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, pp. 531–537, 1999.

[4] 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. Natl Acad. Sci. USA*, vol. 96, pp.6745–6750, 1999.

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

[6] T.K. Paul and H. Iba, “Prediction of cancer class with majority voting genetic programming classifier using gene expression data,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol 6, no.2, pp.353 – 367, 2009.

[7] K. Bryan, P. Cunningham, and N. Bolshakova, “Application of simulated annealing to the biclustering of gene expression data,” *IEEE Trans. Information Technology in Biomedicine*, vol.10, no.3, pp.519-525, 2006.

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

[9] 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.

[10] Y.Tang, Y. Zhang, and Z.Huang, “Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis,” *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol.4, no.3, pp.365-381, 2007.

[11] J. Reunanen, “Overfitting in making comparisons between variable selection methods,” J. Machine Learn. Res., vol.3, pp.1371–1382, 2003.

[12] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” SIAM Review, vol. 43, no. 1, pp. 129–159, 2001.

[13] E. J. Cand`es, J. Romberg, and T. Tao, “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information,” IEEE Transactions on Information Theory, vol. 52, no. 2, pp. 489–509, 2006.

[14] E. J. Cand`es and T. Tao, “Near-optimal signal recovery from random projections: universal encoding strategies?” IEEE Transactions on Information Theory, vol. 52, no. 12, pp. 5406–5425, 2006.

[15] D. L. Donoho, “Compressed sensing,” IEEE Transactions on Information Theory, vol. 52, no. 4, pp. 1289–1306, 2006.

[16] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” Journal of the Royal Statistical Society. Series B, vol. 58, no. 1, pp. 267–288, 1996.

[17] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, “An interior-point method for large-scale l1-regularized least squares,” IEEE Journal on Selected Topics in Signal Processing, vol. 1, no. 4, pp. 606–617, 2007.

[18] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210–227, 2009.

[19] X. Hang, and F.X. Wu, “Sparse Representation for Classification of Tumors Using Gene Expression Data,” Journal of Biomedicine and Biotechnology, vol. 2009, Article ID 403689, 6 pages.

[20] P. Belhumeur, J. Hespanda, and D. Kriegman, “Eigenfaces vs. Fisherfaces: recognition using class specific linear projection,” IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711–720, 1997.

[21] C.H. Zheng, D.S. Huang, L. Zhang, and X.Z. Kong, “Tumor clustering using non-negative matrix factorization with gene selection,” IEEE Transactions on Information Technology in Biomedicine. 13(4), pp.599-607, 2009.

[22] O. Alter, P.O. Brown, and D. Botstein, “Singular value decomposition for genome-wide expression data processing and modeling,” *Proc. Natl. Acad. Sci.*, vol. 97, pp.10101-10106, 2000.

[23] W. Liebermeister, “Linear modes of gene expression determined by independent component analysis,” *Bioinformatics*, vol. 18, pp. 51-60, 2002.

[24] C.H. Zheng, D.S. Huang, K. Li, G. Irwin, and Z.L. Sun, “MISEP method for post-nonlinear blind source separation,” Neural Computation, vol.19, no.9, pp.2557-2578, 2007.

[25] H.Q. Wang, and D.S. Huang, "Regulation probability method for gene selection," Pattern Recognition Letter, vol.27, no.2, pp.116-122, 2006.

[26] H.Q. Wang, H.S.Wong, D.S. Huang, and J. Shu, “Extracting gene regulation information for cancer classification,” Pattern Recognition, vol. 40, pp. 3379-3392, 2007.

[27] E. Amaldi and V. Kann, “On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems,” Theoretical Computer Science, vol. 209, pp. 237–260, 1998.

[28] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, “Support vector machines classification and validation of cancer tissue samples using microarray expression data,” Bioinformatics, vol.16, pp.906–914, 2000.

[29] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, “A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis,” Bioinformatics, vol. 21, no. 5, pp. 631–643, 2005.

[30] N. Pochet, F. De Smet, J.A.K. Suykens, and B.L.R. De Moor, “Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction,” Bioinformatics, vol.20, pp.3185-3195, 2004.

[31] J. E. Staunton, D. K. Slonim, H. A. Coller, et al., “Chemosensitivity prediction by transcriptional profiling,” Proceedings of the National Academy of Sciences of the United States of America, vol. 98, no.19, pp.10787–10792, 2001.

[32] A. I. Su, J. B. Welsh, L. M. Sapinoso, et al., “Molecular classification of human carcinomas by use of gene expression signatures,” Cancer Research, vol. 61, no. 20, pp. 7388–7393, 2001.

[33] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, and S.J. Korsmeyer, “MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia,” Nat. Genet., vol.30, pp.41–47, 2002.

[34] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, et al., “Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses,” Proc. Natl. Acad. Sci. USA **98**, pp. 13790–13795, 2001.

[35] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P.S. Meltzer, “Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks,” *Nature Medicine,* vol. 7, no. 6, pp. 673-679, 2001.

[36] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, et.al., “Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning,” Nature Medicine, Vol. 8, No. 1. pp. 68-74, 2002.

[37] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D’Amico, J.P. Richie, et al., “Gene expression correlates of clinical prostate cancer behavior,” Cancer Cell, vol.1, pp.203–209, 2002.

[38] S. Dudoit, J. Fridlyand, and T.P. Speed, “Comparison of discrimination methods for the classification of tumor using gene expression data,” J. Am. Stat. Assoc., vol. 97, pp. 77–87, 2002.

[39] S.L. Wang, X. Li, S. Zhang, J. Gui, and D.S. Huang, “Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction,” Computers in Biology and Medicine, vol. 40, no.2, pp. 179-189, 2010.

[40] D. Ghosh, and A.M. Chinnaiyan, “Classification and selction of biomarks in genomic data using LASSO,” Journal of Biomedicine and Biotechnology, vol.2, pp.147-154, 2005.