Learning a Weighted Meta-Sample Based Parameter Free Sparse Representation Classification for Microarray Data

Bo Liao*, Yan Jiang, Guanqun Yuan, Wen Zhu, Lijun Cai, Zhi Cao

Key Laboratory for Embedded and Network Computing of Hunan Province, the College of Information Science and Engineering, Hunan University, Changsha Hunan, China

Abstract

Sparse representation classification (SRC) is one of the most promising classification methods for supervised learning. This method can effectively exploit discriminating information by introducing a ℓ_1 regularization terms to the data. With the desirable property of sparisty, SRC is robust to both noise and outliers. In this study, we propose a weighted meta-sample based non-parametric sparse representation classification method for the accurate identification of tumor subtype. The proposed method includes three steps. First, we extract the weighted meta-samples for each sub class from raw data, and the rationality of the weighting strategy is proven mathematically. Second, sparse representation coefficients can be obtained by ℓ_1 regularization of underdetermined linear equations. Thus, data dependent sparsity can be adaptively tuned. A simple characteristic function is eventually utilized to achieve classification. Asymptotic time complexity analysis is applied to our method. Compared with some state-of-the-art classifiers, the proposed method has lower time complexity and more flexibility. Experiments on eight samples of publicly available gene expression profile data show the effectiveness of the proposed method.

Citation: Liao B, Jiang Y, Yuan G, Zhu W, Cai L, et al. (2014) Learning a Weighted Meta-Sample Based Parameter Free Sparse Representation Classification for Microarray Data. PLoS ONE 9(8): e104314. doi:10.1371/journal.pone.0104314

Editor: Enrique Hernandez-Lemus, National Institute of Genomic Medicine, Mexico

Received February 27, 2014; **Accepted** July 7, 2014; **Published** August 12, 2014

Funding: This work is supported by the Program for New Century Excellent Talents in University (Grant NCET-10-0365), National Nature Science Foundation of China (Grant 60973082, 11171369, 61272395, 61370171, 61300128), the National Nature Science Foundation of Hunan province (Grant 12JJ2041), the Planned Science and Technology Project of Hunan Province (Grant 2009FJ3195, 2012FJ2012) and supported by the Fundamental Research Funds for the Central Universities, Hunan university, Graduate research and innovation projects in Hunan Province (Grant CX2012B144). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: dragonbw@163.com

Introduction

The development of high-throughput technologies has enabled scientists to monitor the gene expression levels in tens of thousands of genes simultaneously in a single experiment. This technology has become a symbol of the post-genomic era [1]. Biomedical research indicates that tumor development is related to the change in gene expression levels and that tumor-related biomarkers are usually associated with a few genes. Thus, identifying tumor tissue or disease-related biomarkers accurately is of great practical significance. However, gene expression profile data are characterized by very high dimensionalities and small sample size. The curse of dimensionality problem makes classification challenging.

Some dimensionality reduction methods have recently been proposed to solve the "large *p*, small *n*" problem [2]. Feature extraction and feature selection are two methods of dimensionality reduction; feature extraction transforms original features (genes) into a set of new features by subspace learning [3–5]. However, suitable biological interpretation is difficult to obtain from the subspace learning dimensionality reduction results. Feature selection is another commonly used dimensionality reduction method that selects a sub-set of genes that can best predict the response values from the raw data [6]. Although dimensionality reduction can significantly improve computational efficiency, this process can easily lead to over-fitting when a classifier is applied.

Sparse representation classification (SRC) was proposed by Wright et al. [7] for face recognition. With ℓ_1 sparsity constraint, a testing face can be approximately represented by parts of the training data that are from the same class. Unlike traditional classification methods such as support vector machine and *k* nearest neighbor classifier, SRC is robust to both noise and outliers. However, the orginal training samples may not contain suffiient discriminating information compared with meta-samples [8].

To capture more alternative information from gene expression data, the so-called meta-samples are proposed by [8–11]. These samples can be regarded as a set of bases, the linear representation of which can represent the training data. In [11], penalized matrix decomposition is used to extract meta-samples, and clustering is performed on those meta-samples. In [8], the meta-sample based sparse representation classification (MSRC) method is proposed. This method is robust to over-fitting problem and noise. However, MSRC needs two predefined parameters, namely, the number of meta-samples and the sparse penalty factor. These two parameters are data dependent. Thus, model selection methods, such as cross-validation (CV), significantly affect the classification results. In this study, we propose a non-parametric version of MSRC to address this optimal parameter selection problem. The main contributions of this paper are as follows:

Table 1. Notations and abbreviations used in this paper.

Notation	Description
SVD	Singular value decomposition
$\Re^N$	N dimensional real number vector
$\mathcal{X}$	$\mathcal{X}=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_m\}\in\Re^{n\times m}$ denotes gene expression data set with n genes, m samples
$\mathbf{W}$	$\mathbf{W}=[\mathbf{W}_1,\mathbf{W}_2,...,\mathbf{W}_c]$ meta-samples associate with c classes
$\lvert c_i\rvert$	Number of samples belong to class i
$\|\|_0$	ℓ_0 norm
$\|\|_1$	ℓ_1 norm
$\|\|_F$	Matrix Frobenius norm

doi:10.1371/journal.pone.0104314.t001

1. The data-dependent sparsity can be automatically adjusted, rather than empirically chosen. Without computationally expensive model selection, our method is scalable and efficient.
2. The existing MSRC [8] method requires the appropriate selection of the number of meta-samples for each sub class, which is a laborious task. We address this problem by introducing a simple weighting strategy for the meta-sample of each category, and the rationality of weighting strategies is mathematically proved.
3. Extensive experiments are performed to evaluate the proposed method. Experimental results show the superiority of the non-parametric version of MSRC compared with some state-of-the-art classifiers. Section 3 presents more details.

The remainder of this paper is organized as follows: prior work on sparse representation classification and the fundamentals of the proposed method are described in Section 2. Section 3 presents the experimental results. The proposed method is discussed in Section 4. Section 5 concludes this paper.

Methods

This study primarily aims to establish the manner by which to devise an robust classifier for tumor subtype classification. Given a microarray data set $\mathcal{X}=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_m\}\in\Re^{n\times m}$ and a set of class labels $C=\{1,2,...,c\}$, $\mathbf{X}$ is a matrix with n rows and m columns. Each column of $\mathbf{X}$ denotes a sample, whereas each row of $\mathbf{X}$ denotes a gene. Let $\mathbf{x}_j$ denote the jth sample, which is a column vector with n dimensional. For each element in $\mathbf{X}$, $x_{i,j}\in\Re$ denotes the expression level of the ith gene in the jth sample. We provide a summary of the abbreviations used in this study in Table 1. For clarity, we use boldface and lowercase type letters for vectors and boldface and capital type letters for matrices.

Gene expression profile data are high-throughput data with tens of thousands of genes. However, the number of samples is usually very small, which makes classification challenging. To avoid the curse of dimensionality, differential gene expression analysis [12,13] is widely used to exclude redundant and irrelevant genes before classification. In our study, we use the Relieff [14] method to select a subset of informative genes for further analysis. In the following subsections, we briefly review meta-sample and sparse

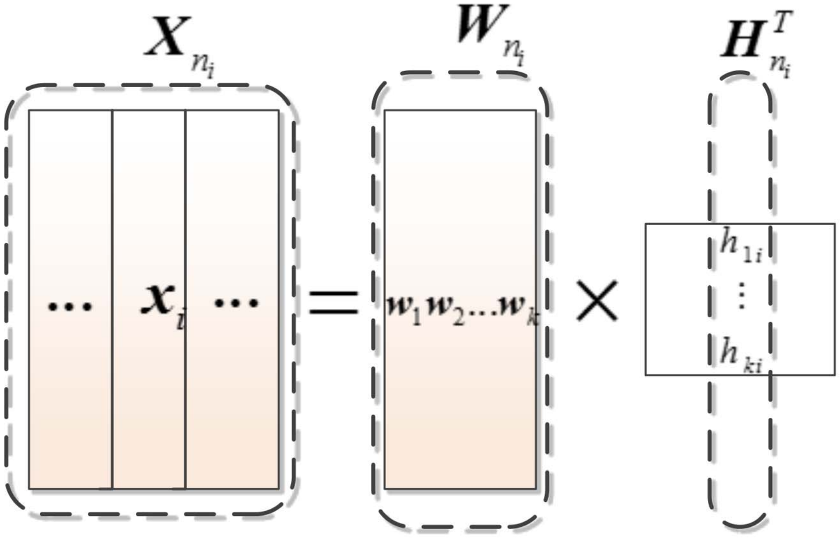

Figure 1. Illustration of meta-sample model: each column vector of $\mathbf{X}_{n_i}$ can be represented within a linear combination of meta-samples in $\mathbf{W}_{n_i}$, and the column of $\mathbf{H}^T_{n_i}$ corresponds to the linear combination coefficients.

doi:10.1371/journal.pone.0104314.g001

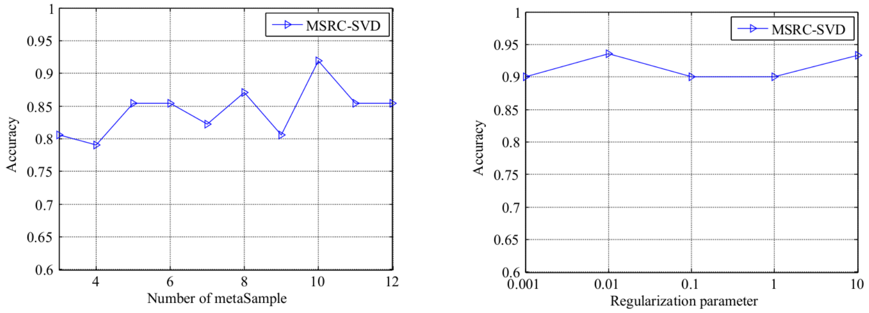

Figure 2. Optimal classification accuracy of MSRC achieved on COLON; the x-axis represents the number of meta-samples (left) and the regularization parameter (right). Classification accuracy is more sensitive to the number of meta-samples rather than to the regularization parameter.
doi:10.1371/journal.pone.0104314.g002

representation classification. we then propose weighted meta-sample based parameter free sparse representation classification (PFMSCR).

Meta-samples versus gene expression samples

As illustrated in Figure 1, meta-samples can be regarded as basis samples that contain the essential information of the original data. A given testing sample can be represented by a linear combination of meta-samples from the same class. Concretely, suppose $\mathbf{x}_i$ is associated with the $n_i th$ class, where $n_i \in C$, and the $n_i th$ class samples in the training data have k meta-samples, namely, $\{\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_k\} \in \Re^{n\times k}$. Sample $\mathbf{x}_i$ can be formulated as Eq. (1).

$$\mathbf{x}_i = \mathbf{w}_1 h_{1,i} + \mathbf{w}_2 h_{2,i} + \ldots + \mathbf{w}_k h_{k,i} \tag{1}$$

Mathematically, meta-samples extraction can be regarded as a type of matrix decomposition, including non-negative matrix factorization [15], singular value decomposition (SVD) [16], and principal component analysis [17], where matrix $\mathbf{W}_{n_i} \in \Re^{n\times k}$, and $\mathbf{H}_{n_i}^T \in \Re^{k\times|n_i|}$ denote the meta-sample and meta-gene, respectively. In singular value decomposition, $\mathbf{W}_{n_i}$ is a maximum linearly independent group of $\mathbf{X}_{n_i}$ column vectors.

Biologically, meta-samples are also called eigenarray [18] or basis snapshot for gene expression data. Han et al. [17] used meta-samples to identify tumors from microarray data and found that meta-sample-based classification can effectively avoid over-fitting. Zheng et al. [10,11,18] proposed a novel cluster method based on meta-samples, which meta-samples can be regarded as cluster indictors.

Prior works revealed that meta-samples preserve some desired discriminant information of samples from the same class.

Sparse representation classification problem revisited

In this subsection, we revisit the sparse representation problem briefly. Sparse representation is one of the most important components of machine learning and data mining community that has wide applications in such fields as text mining, image classification, and bioinformatics. In this work, we interpret the sparse representation problem from the view of linear algebra.

From the standpoint of linear equations system $\mathbf{X}\boldsymbol{\alpha}=\mathbf{y}$, the solution of $\mathbf{X}\boldsymbol{\alpha}=\mathbf{y}$ has three possible states:

1. Linear equation systems have infinitely many solutions if they are underdetermined (i.e., $n<m$).
2. Linear equation systems have a unique solution if they are well posed.
3. Linear equation systems have no solution if overdetermined (i.e., $n>m$).

In the first scenario, one can pursue the sparse solution by regularization [19]. The problem can be formulated as

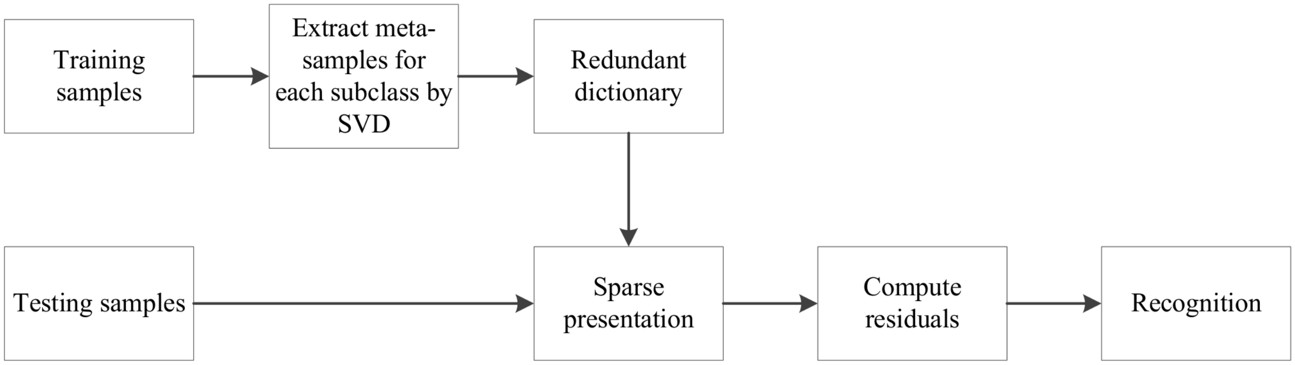

Figure 3. The flowchart of PFMSRC scheme.
doi:10.1371/journal.pone.0104314.g003

Table 2. Descriptions of microarray data repository and the accession number.

Datasets	Repository	Accession number
Colon	Gene Expression Omnibus	GDS4379
Acute leukemia data	Gene Expression Omnibus	GSE19475
DLBCL	Gene Expression Omnibus	GSE15177
Gliomas	Gene Expression Omnibus	GSE54792
SRBCT	Gene Expression Omnibus	GSE1825,GSE31186,GSE31217
ALL	Gene Expression Omnibus	GSE23024
MLLLeukemia	Gene Expression Omnibus	GSE11038
LukemiaGloub	Gene Expression Omnibus	GSE10283

doi:10.1371/journal.pone.0104314.t002

$$\begin{aligned} \min_{\alpha} \quad & \|\boldsymbol{\alpha}\|_0 \\ s.t. \quad & \mathbf{X}\boldsymbol{\alpha} = \mathbf{y} \end{aligned} \tag{2}$$

However, ℓ_0 norm is an NP-hard combinational optimization problem, and difficult to solve, fortunately, ℓ_1 norm is an appropriate convex approximate to ℓ_0 [20]. If the solution is sparse enough, ℓ_1 minimization is equivalent to ℓ_0 minimization [21], such that we can reformulate Eq. (2) as

$$\begin{aligned} \min_{\alpha} \quad & \|\boldsymbol{\alpha}\|_1 \\ s.t. \quad & \mathbf{X}\boldsymbol{\alpha} = \mathbf{y} \end{aligned} \tag{3}$$

For the other two scenarios, the sparsity of $\boldsymbol{\alpha}$ cannot be guaranteed. However, one can still obtain a sparse solution by adding a penalty term that shares the same formulation as LASSO [22]

$$\min_{\alpha} \quad \|\mathbf{X}\boldsymbol{\alpha} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \tag{4}$$

Compared with Eq. (3), Eq. (4) is an unconstrained convex problem. Notably, λ makes a tradeoff between sparsity and regression error and should be empirically chosen. A larger λ yields a sparser α. However, one might run the risk of increasing regression error term $\|\mathbf{X}\boldsymbol{\alpha} - \mathbf{y}\|_2^2$.

Sparse representation assumes that a signal can be reconstructed by a small number of basis signals within a linear combination. Thus, Eq (3) can be named as basis pursuit [23]. In bioinformatics applications, one can suppose that a testing sample can be well reconstructed by the training data from the same class within a linear combination, which is a very useful assumption for our later work.

Meta-sample based sparse representation

Zheng et al. [8] proposed MSRC method to predict tumor subtypes. In such situations, c classes of meta-samples are extracted, denoting as $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, ..., \mathbf{W}_c]$ with the same classes being conjoined together, where meta-samples are column vectors (two kinds of meta-sample are proposed in [8]). Given a test sample $\mathbf{y}$ associated with class i, MSRC tries to find sparse reconstruct coefficients in terms of all meta-samples using Eq. (4). In particular, [8] tries to solve the sparse representation problem using $\min_{\alpha}\|\mathbf{W}\boldsymbol{\alpha} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$. In ideal cases, the nonzero entries in α will only be associated with the ith class meta-samples of $\mathbf{W}$, as shown in Eq. (5).

$$\boldsymbol{\alpha} = [\mathbf{0}, ..., \underbrace{\alpha_{i1}, \alpha_{i2}, ..., \alpha_{in_i}}_{ith\ class}, ..., \mathbf{0}]^T \in \Re^m \tag{5}$$

Notably, the gene expression profile contains data with high dimensionality and small sample size ($n \gg m$). The sparsity can only be achieved by adding a penalty term. However, the optimal number of meta-samples and penalty factor λ are essentially important in classification applications. Figure 2 illustrates that if the meta-samples are improperly set, the prediction accuracy of MSRC drops seriously on COLON dataset. Specifically, in the left part of Figure 2 shows that the 10-fold stratified cross validation classification accuracy is achieved by varying the number of meta-samples from 3 to 12 for each subclass. We can observe that the performance is less sensitive to various regularization parameters within the scope of λ from the right part of Figure 2. Thus, model selection is essential and laborious work on different data sets.

To overcome this weakness, this study proposed a novel parameter free meta-sample based sparse representation classification (PFMSRC) method.

Parameter free meta-sample sparse representation (PFMSRC)

In this subsection, we first propose a heuristic weighted strategy, the reasonableness of which is theoretically proven. We then construct an underdetermined linear equation system, in which the data-dependent sparsity can be self-adaptively tuned by ℓ_1 norm regularizer.

Let $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_c\} \in \Re^{n \times m}$ be gene expression profile data, with the same classes being conjoined together, that is, $\mathbf{X}_i$ contains all samples associated with the ith class. We factorize $\mathbf{X}_i$ by performing SVD. The singular values are sorted in descending order $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k > 0$, where k is the column rank of $\mathbf{X}_i$, and $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, ...\lambda_k)$ denotes diagonal matrix with singular values being diagonal elements. One can extract weighted meta-samples associated with class i as $\mathbf{W}_i = [\sqrt{\lambda_1}\mathbf{u}_1, \sqrt{\lambda_2}\mathbf{u}_2, ..., \sqrt{\lambda_k}\mathbf{u}_k]$, where u_i is a column vector in U_i, and $rank(\mathbf{X}_i) = k$.

Table 3. Data set descriptions.

Datasets	Samples	Genes	Subclass number
Colon	62	2000	2
Acute leukemia data	72	5000	2
DLBC	77	7129	2
Gliomas	50	12625	2
SRBCT	83	2308	4
ALL	248	12626	6
MLLLeukemia	72	12582	3
LukemiaGloub	72	7129	3

doi:10.1371/journal.pone.0104314.t003

$$\mathbf{X}_i=\mathbf{U}_i\begin{pmatrix}\lambda_1 & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots\\ \mathbf{0} & \cdots & \lambda_k\end{pmatrix}\mathbf{V}_i^{\ T},\quad \forall i,\lambda_i>0 \qquad (6)$$

Alternatively, Eq. (6) can be compactly reformulated as $\mathbf{X}_i=\mathbf{U}_i\sqrt{\mathbf{\Lambda}}\sqrt{\mathbf{\Lambda}}\mathbf{V}_i^T$. This weighting scheme can enhance the influence of main singular vector in $\mathbf{U}_i$. That is, larger λ_i makes the associated meta-sample more important. Moreover, the weighting scheme works well in the following experiments. Compared with [8], Zheng et al. extracted meta-samples by performing SVD as well. However, in their algorithm framework, the number of meta-samples used for classification is determined during the cross-validation step. On the contrary, PFMSRC tries to avoid the cross-validation part by weighting the all meta-samples and weakening the influence of minor eigenvectors rather than using several of them for classification. Proposition 1 theoretically proves the reasonableness of the weighting strategy in measuring the importance of each metasample.

Proposition 1. *Singular value is a reasonable weighting factor for measuring the importance of meta-samples.*

Proof. Let $\mathbf{X}=[u_1,u_2,...,u_k]\mathbf{\Lambda}[v_1,v_2,...,v_k]^T$, where $\mathbf{\Lambda}=diag(\lambda_1,\lambda_2,...,\lambda_k)$ and $\lambda_1\geq\lambda_2\geq...\geq\lambda_k>0$, $\mathbf{X}=\sum_{i=1}^{k}\lambda_i\mathbf{u}_i\mathbf{v}_i^T$, considering evaluation metric function $\frac{\|\lambda_i\mathbf{u_i}\mathbf{v}_i^T\|_F^2}{\|\mathbf{X}\|_F^2}=\frac{Tr(\lambda_i^2\mathbf{u_i}\mathbf{v}_i^T\mathbf{v_i}\mathbf{u}_i^T)}{Tr(\mathbf{XX}^T)}=\frac{\lambda_i^2}{\lambda_1^2+\lambda_2^2+...+\lambda_k^2}\geq\frac{\lambda_j^2}{\lambda_1^2+\lambda_2^2+...+\lambda_k^2}$, one can conclude that

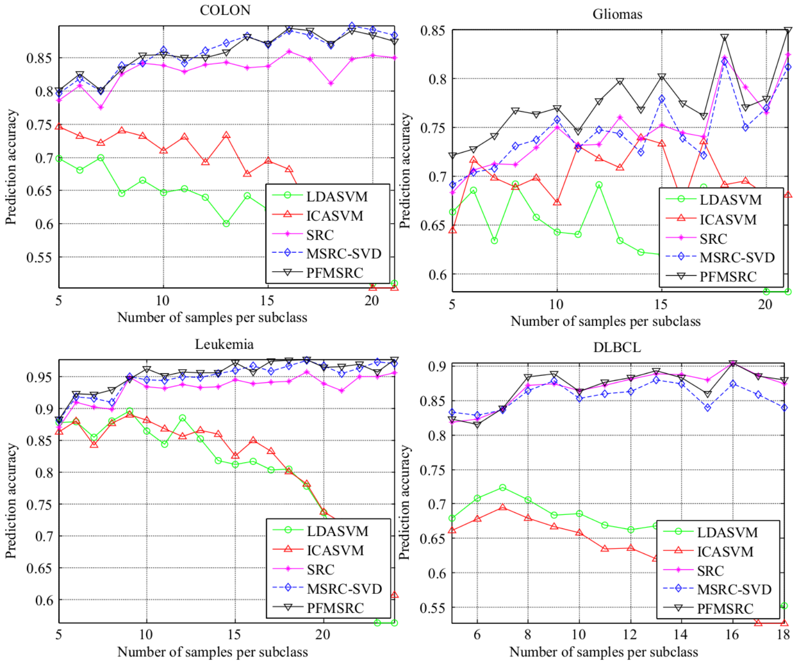

Figure 4. Comparison of prediction accuracy on four binary classification datasets by varying the number of samples from per subclass; when p is larger than 10 the model based method prediction accuracy decreases as p increases.
doi:10.1371/journal.pone.0104314.g004

Table 4. Comparison on four binary classification tumor data sets; for each data set, 10 samples per class are randomly selected for training and the rest are used for testing.

Dataset name	LDA+SVM	ICA+SVM	SRC	MSRC-SVD	PFMSRC
colon	74(±7.85)	64.55(±7.39)	84.20(±3.65)	84.20(±4.81)	**85.45(±3.33)**
DLBC	66.76(±6.67)	68.33(±4.78)	**86.49(±3.39)**	85.35(±4.91)	86.40(±5.69)
Gliomas	65.83(±8.08)	69.83(±9.52)	75.00(±6.35)	75.83(±7.24)	**77.00(±6.48)**
Acute leukemia	89.71(±3.14)	89.13(±4.96)	93.46(±3.82)	94.52(±3.65)	**96.25(±2.20)**

We report the standard deviations in parentheses.

doi:10.1371/journal.pone.0104314.t004

$$\frac{\|\lambda_i \mathbf{u_i} \mathbf{v}_i^T\|_F^2}{\|\mathbf{X}\|_F^2} \geq \frac{\left\|\lambda_j \mathbf{u}_j \mathbf{v}_j^T\right\|_F^2}{\|\mathbf{X}\|_F^2}$$

This completes the proof. □

The evaluation metric function is used to measure the meta-sample's contribution of the meta-sample to the raw data reconstruction in terms of λ_i. Tr denotes matrix trace. Note that, functions $f(x)=x$ and $g(x)=\sqrt{x}$ have the same monotonicity, which makes the weighting strategy reasonable.

ℓ_1 graph was proposed by Cheng et al. [24] to measure the similarity among samples. Inspired by their work, sparsity can be obtained by ℓ_1 regularizer on underdetermined linear equation systems. Concretely, a testing sample can be recovered by weighted meta-samples within a linear combination with a noise term added, formulated as Eq. (7)

$$\mathbf{y}=\mathbf{W}\boldsymbol{\alpha}+\mathbf{e}=[\mathbf{W}\ \mathbf{I}]\begin{bmatrix}\boldsymbol{\alpha}\\ \mathbf{e}\end{bmatrix} \tag{7}$$

Let $\mathbf{B}=[\mathbf{W}\ \mathbf{I}]\in\Re^{n\times(m'+n)}$ and $\boldsymbol{\alpha}'=\begin{bmatrix}\boldsymbol{\alpha}\\ \mathbf{e}\end{bmatrix}\in\Re^{m'+n}$, where m' represents the number of meta-samples corresponding to c classes, I is an identity matrix, and $\mathbf{e}$ is the noise term. Alternatively, one can solve the following minimization problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha}'} \quad & \|\boldsymbol{\alpha}'\|_1 \\ s.t. \quad & \mathbf{y}=\mathbf{B}\boldsymbol{\alpha}' \end{aligned} \tag{8}$$

Theorem 1 proves that Eq. (8) is a underdetermined linear system. As stated in Subsection 2.2 the sparsity of underdetermined linear system can be automatically tuned by ℓ_1 regularization (the first scenario). Moreover, (8) is a canonical convex problem with equality constraints, which can optimize sparse representation coefficients and noise term simultaneously. The globally optimal solution can be efficiently solved by CVX package [25] in polynomial time. Notably, the package solves the optimization problem by dualization rather than interior point method because the former is significantly faster than the latter.

Theorem 1. *Linear equation system (8) is underdetermined, and* $rank(\boldsymbol{B})=n$.

Proof. We can find a sub matrix in $B\in\Re^{n\times(m'+n)}$, such as $\mathbf{I}$ and $rank(\mathbf{I})=n\Rightarrow rank(\mathbf{B})=n$. This completes the proof. □

Note that $\boldsymbol{\alpha}'\in\Re^{m'+n}$ is a sparse vector with $m'+n$ entries. The first m' components correspond to linear representation coefficients, whereas the last n components characterize model noise or regression error. However, the test sample $\mathbf{y}$ from one of the classes in training data cannot be well reconstructed by meta-samples associated with the same class in most instances because of the existence of noises. Figure 3 illustrates the flowchart of our PFMSCR scheme, the redundant dictionary is constructed by combining meta-samples and noise term.

We define a projection function $\delta_i(\boldsymbol{\alpha}'):\Re^{m'}\rightarrow\Re^{m'}$ for each class i, which selects the coefficients associated with the ith class from the first m' components in $\boldsymbol{\alpha}'$, whereas the other entries are appropriately padded with zeros in $\delta_i(\boldsymbol{\alpha}')$. The reconstruction relationship $\mathbf{y}=\mathbf{W}\delta_i(\boldsymbol{\alpha}')$ is not always holden. However, the minimized reconstruction error criterion $r_i(\mathbf{y})=\|\mathbf{y}-\mathbf{W}\delta_i(\boldsymbol{\alpha}')\|_2$, $i=1...c$ is a good approximation to classify testing samples. We summarize the proposed classification method as follows.

Step 1. Input training sets $\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_c]\in\Re^{n\times m}$, class number c, and testing sample $\mathbf{y}\in\Re^n$;

Step 2. Normalize training set samples and testing sample to obtain unit ℓ_2-norm;

Step 3. Extract weighted meta-samples $\mathbf{W}=[\mathbf{W}_1,\mathbf{W}_2,...,\mathbf{W}_c]$ for each class (meta-samples with the same class are conjoint);

Step 4. Solve non-parametric sparse representation problem by Eq. (8);

Table 5. Comparison of specificity by different methods on four binary classification data sets.

Dataset name	SRC	MSRC-SVD	PFMSRC
colon	90.00	**92.50**	**92.50**
DLBC	**96.55**	94.83	**96.55**
Gliomas	72.73	**77.27**	**77.27**
Acute leukemia	100	100	100

doi:10.1371/journal.pone.0104314.t005

Table 6. Comparison of sensitivity by different methods on four binary classification data sets.

Dataset name	SRC	MSRC-SVD	PFMSRC
colon	81.82	**86.36**	**86.36**
DLBC	**1**	**1**	94.74
Gliomas	82.14	78.57	**89.29**
Acute leukemia	88.00	**92.00**	84.00

doi:10.1371/journal.pone.0104314.t006

Step 5. Compute residuals for each class $r_i(\mathbf{y}) = \|\mathbf{y}-\mathbf{W}\delta_i(\boldsymbol{\alpha}')\|_2,\ \ i=1\ldots c$;

Step 6. Return class label of $\mathbf{y}$ as $c(\mathbf{y}) = \arg\min_i r(\mathbf{y}),\ \ i=1,\ldots,c$;

PFMSRC can be considered as a non-parametric version of MSRC, compared with the former having the following merits:

1. The weighted meta-samples are orthogonal with one another. That is, no redundancy exists among meta-samples, and the weight enhances the influence of the main singular vector, such that discriminant information can be well retained.
2. The data-dependent sparsity can be automatically tuned without human intervention. Thus, PFMSRC has better scalability and robustness.
3. The time complexity of PFMSRC is lower than that of MSRC, since computationally expensive model selection work need not be accomplished for parameter optimization. Time complexity can be estimated as: weighted meta-sample extraction step needs time complexity $\mathcal{O}(nm^2)$, ℓ_1 minimization needs time complexity $\mathcal{O}((m+n)^{1.3})$, the total complexity for PFMSRC is $\mathcal{O}(nm^2+m(m+n)^{1.3})$.

In the following section, we will conduct extensive experiments on micoarray data to evaluate the effectiveness of our scheme, and microarray data repository information as well as the accession number is given by Table 2.

Experiments

In this section, we will evaluate the performance of the proposed PFMSRC algorithm against four state-of-the-art algorithms, namely, linear discriminant analysis (LDA+SVM), independent component analysis (ICA+SVM), SRC, and meta-sample sparse representation (SVD-MSRC). The former two are model based and accompanied by feature extraction. These two algorithms are regarded as baseline. For the model-based method, support vector machine [26,27] with radial basis function kernel is employed as a classifier. The experiments are performed on four binary-class classification data sets and four multiclass classification data sets. All experiments are implemented in Matlab environment and run on a personal computer with intel Pentium4 dual core CPU 2.4 GHZ and 4 G RAM. The summarized descriptions of the eight gene expression profile datasets are provided by Table 3.

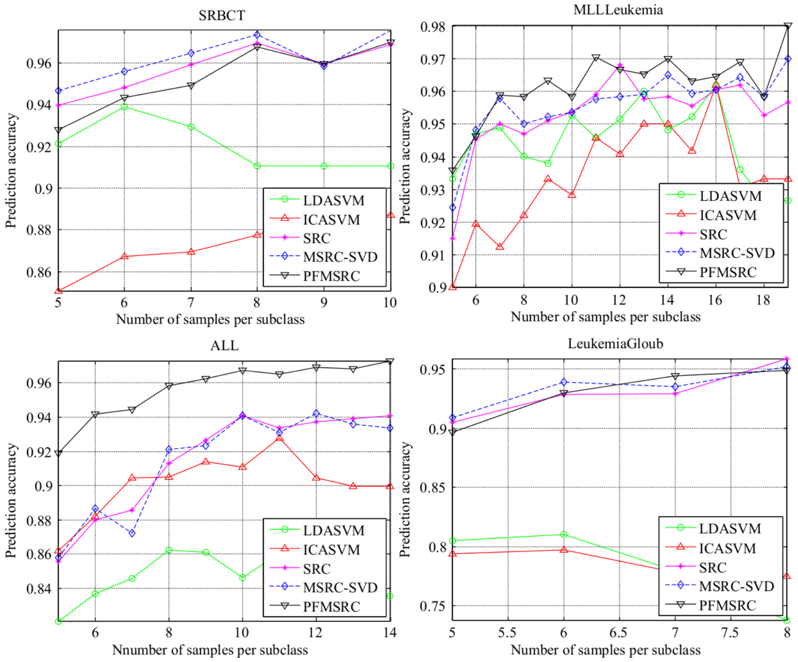

Figure 5. Comparison of prediction accuracy on four multiclass classification datasets by varying the number of samples from per subclass; when p is larger than 10 the performance degradation of model based methods is less significant than that of binary classification.
doi:10.1371/journal.pone.0104314.g005

Table 7. Comparison on four multiclass tumor data sets; for each data set, 10 (8 for LeukemiaGloub) samples per class are randomly selected for training the rest are used for testing.

Dataset name	LDA+SVM	ICA+SVM	SRC	MSRC-SVD	PFMSRC
SRBCT	91.05(±4.61)	88.72(±5.56)	96.86(±2.64)	**97.56(±3.06)**	96.98(±2.51)
ALL	86.12(±3.81)	91.38(±3.28)	94.07(±2.38)	94.07(±2.93)	**96.73(±1.68)**
MLLLeukemia	93.81(±3.74)	93.33(±5.16)	95.36(±3.04)	95.36(±2.84)	**95.83(±2.88)**
LukemiaGloub	73.75(±5.25)	77.50(±6.98)	**95.83(±2.14)**	95.21(±2.35)	94.90(±2.74)

The average accuracy and corresponding standard deviations are reported.
doi:10.1371/journal.pone.0104314.t007

- Colon [28] consists of 62 samples with two subclasses including 40 tumor and 22 normal samples. The highest 2000 genes with minimal intensity in the tissues are retained from the original of more than 6500 genes. This dataset can be downloaded from [29].
- Acute leukemia data [30], consist of 72 samples with two subclasses, including 47 acute lymphoblastic leukemia patients and 25 acute myelogenous leukemia patients. Each sample contains 7129 genes. This dataset can be downloaded from [29].
- DLBCL [1] consists of 77 samples with two subclasses, including 58 diffuse large b-cell lymphoma samples and 19 follicular lymphoma samples. Each sample contains 7129 genes. This dataset can be downloaded from [31].
- Gliomas [32] consist of 50 samples with two subclasses (Glioblastomas and Anaplastic Oligodendrogliomas), and each sample contains 2308 genes. This dataset are available at [31].
- SRBCT [33] consist of 83 samples with four subclases (Ewings sarcoma, Burkitts, Neuroblastoma and rhabdomyosarcoma). Each sample contains 2308 genes. The datasets are available at [31]
- ALL [34] consists of 248 samples with six subclasses. Each sample contains 12626 genes. The datasets are available at [31].
- MLLLeukemia [35] consists of 72 samples with three subclasses. Each sample contains 12582 genes. The datasets are available at [29].
- LukemiaGloub [30] consists of 72 samples with three subclasses. Each sample contains 7129 genes. The datasets are available at [31].

Dataset preprocessing and experiment setup

Gene expression profiling involves data with high dimensionality and small sample size. The exclusion of redundant and irrelevant data is critical for classification. As suggested by [36], restaining only the top 400 genes makes a good tradeoff between computational complexity and biological significance. In our experiment, the top 400 genes are selected from each dataset by applying the Relieff [14] algorithm to the training set.

For LDA+SVM algorithm, we simply extract $c-1$ new features to train the classifier, as LDA can find at most $c-1$ meaningful projection vectors in the subspace, where c denotes the number of

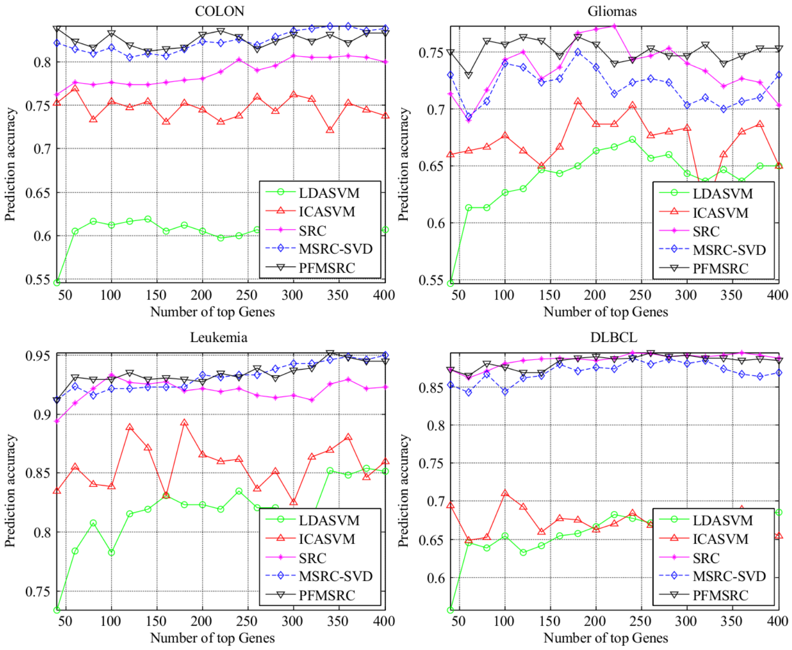

Figure 6. Comparison of prediction accuracy on four binary classification datasets by varying the number of top selected genes.
doi:10.1371/journal.pone.0104314.g006

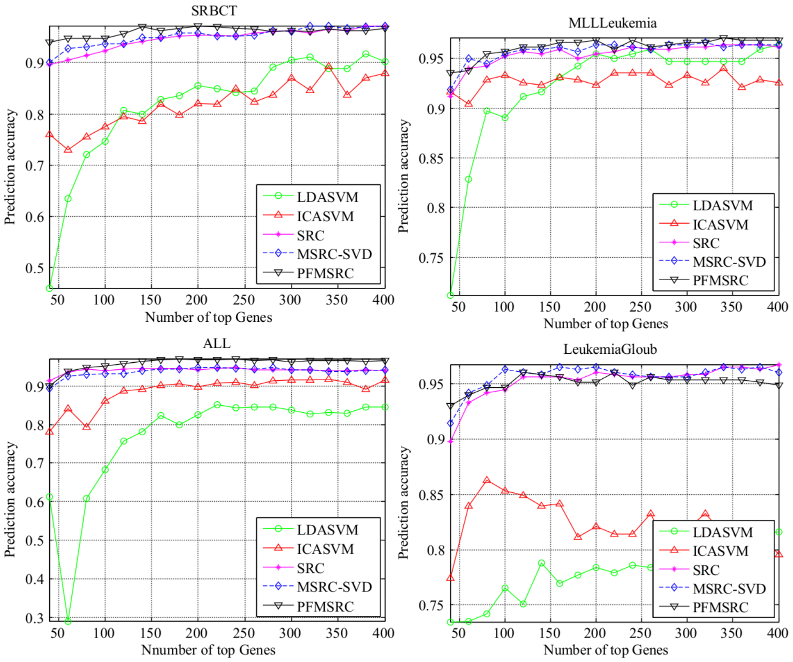

Figure 7. Comparison of prediction accuracy on four multiclass classification datasets by varying the number of top selected genes.
doi:10.1371/journal.pone.0104314.g007

classes. SVM kernel parameters are determined by 10-fold cross-validation. In fact, the determination of the number of independent components is also an empirically dependent work. Here, we use the same method as suggested by [18].

SRC and MSRC methods need parameter λ to control sparsity. MSRC also needs the number of meta-samples of each class as a key parameter. Each dataset λ is searched from $\{0.001,0.1,1,10,100\}$ by 10-fold CV on training data, and the number of meta-samples for each class is set as recommended by [8].

Experiments on binary classification problem

To evaluate the performance of five methods on a balanced split data set, we randomly select $p=5$ to $\min(|c_i|)-1$ samples per subclass as training set and use the rest for testing to guarantee that at least one sample in each category can be used for test, 20 times training/testing are randomly split, and the average classification accuracies are presented. The best prediction accuracy is in boldface for each gene expression profile dataset.

We show the average performance comparison on four binary classification tasks in Figure 4. PFMSRC exhibited encouraging performance. Although Gliomas was difficult for classification, the proposed approach can still achieve 85% classification accuracy via 20 samples per subclass used for training. Notably, the classification accuracy of LDA+SVM and ICA+SVM dropped quickly as more samples are taken for training; the same observations can be found in [36]. This fluctuation phenomenon can be interpreted as follows: (1) For the binary classification case, the feature extracted by LDA has only one dimension that is insufficient to capture the intrinsic discriminating information. Thus, model-based classification methods have difficulty in preventing the over-fitting phenomenon. (2) When evaluating the performance on the testing set the number of samples changes as more samples are used for training.

Table 8. The maximal average prediction accuracy of LDA+SVM, ICA+SVM, SRC, MSRC-SVD and PFMSRC on eight tumor microarray datasets.

Dataset name	LDA+SVM	ICA+SVM	SRC	MSRC-SVD	PFMSRC
colon	61.67	76.90	80.48	**84.05**	83.81
DLBC	68.07	71.05	**89.47**	88.42	89.47
Gliomas	67.33	70.67	75.33	75.00	**76.00**
Acute leukemia	85.38	88.85	93.27	95.00	**95.19**
SRBCT	91.16	89.30	**97.21**	**97.21**	**97.21**
ALL	85.16	91.44	96.46	93.59	**97.02**
MLLLeukemia	96.43	94.05	96.43	96.67	**97.14**
LukemiaGloub	81.63	91.81	94.79	94.68	**96.05**

doi:10.1371/journal.pone.0104314.t008

Table 9. 10-fold CV prediction accuracy of eight tumor microarray datasets using different classification methods.

Dataset name	LDA+SVM	ICA+SVM	SRC	MSRC-SVD	PFMSRC
colon	81.67	90.00	87.14	**90.24**	**90.24**
DLBCL	92.14	**97.14**	97.14	91.96	95.89
Gliomas	**86.50**	**86.50**	78.33	78.33	84.00
Acute leukemia	96.50	95.57	96.07	**97.50**	95.00
SRBCT	96.64	95.75	**1**	**1**	**1**
ALL	97.61	94.83	96.46	93.59	**97.63**
MLLLeukemia	95.65	95.89	**98.75**	**98.75**	97.32
LukemiaGloub	97.32	96.32	**98.57**	**98.57**	96.07

doi:10.1371/journal.pone.0104314.t009

Classification accuracy, specificity, and sensitivity are some popular evaluation metrics. In this work, we use all three to evaluate performance, and the results are reported in Table 4, 5, and 6, respectively. The three methods can achieve satisfactory performance not only on the specificity metric but also on the sensitivity metric. Compared with SRC and MSCR, PFMSRC outperforms its competitors in most cases. A comprehensive consideration is that PFMSRC achieves the best performance, followed by MSRC and SRC.

Experiments on multiclass classification problem

We investigate multiclass classification performance on four publicly available data sets. The experimental setup is the same as that for the binary classification case. On one hand from Figure 5 and Table 7 it can be seen that (1) the classification accuracies of SRC, MSRC, and PFMSRC are increased on all multiclass classification datasets as more samples per subclass are taken for training. (2) ALL has six subclasses, and the proposed PFMSRC achieves the highest classification accuracy, which indicates that we have potential superiority on multiclass classification task. (3) LDA can capture more discriminating information on the multiclass classification task, and the over-fitting phenomenon is reduced compared with the binary classification task.

On the other hand, sparse representation based classification methods are less sensitive to the number of samples used for training model-based classification methods, which suggests a natural approach to select a classifier when the training sample size is small. Table 7 provides the performance description of the five classification methods. The proposed PFMSRC method performs consistently well with small standard deviations. On the SRBCT and ALL datasets, PFMSRC achieved 96.98% and 96.73%, respectively.

Experiments with different number of genes

In this subsection, we evaluate the performance of the five methods with different feature dimensions on eight tumor data sets. For the training data, 10 samples per subclass are randomly selected, whereas the remaining samples are used for test. We perform the test with various numbers of genes, starting from 50 to 400 genes in steps of 20. The comparison experiment was performed 20 times, and the average prediction accuracy of our experiments on eight gene expression profile datasets was recorded for evaluation.

The balanced training sets for each dataset ensure fair evaluation as stated by [36]. The experimental result in Figure 6 shows that the proposed PFMRSC performs well when only 100 genes are used. We can observe the similar results in the multi-classification case as well.

In binary classification case, SRC, MSRC, and PFMSRC share the same curve trend. Compared with SRC and MSRC, PFMSRC performs well by using a smaller number of genes, SRC and MSRC can achieve comparable accuracy by using more genes. Evidently, SRC, MSRC, and PFMSRC consistently outperform LDA+SVM and ICA+SVM in all datasets.

In the multiclass classification case, the performance of MSRC, SRC, and PFMSRC is very stable with respect to the number of genes, and all these methods converge fast to the optimal classification rate point. Figure 7 shows that compared with their performance in the binary classification case, SRC, MSRC, and PFMSRC are less influenced by gene dimension. Note that ALL is a multiclass dataset with six subclasses, but PFMSRC can still achieve a higher classification rate of 97% accuracy compared with SRC and MSRC. The same conclusion can be drawn for the SRBCT dataset.

In Table 8, we report the detailed classification accuracy. PFMSRC outperforms its competitors on most gene expression profile datasets, whereas SRC and MSRC-SVD perform the second best.

Comparsion of CV performance

To evaluate the classification performance on imbalanced split training/testing sets, we perform 10-fold stratified CV on tumor subtype dataset. All samples are randomly divided into 10 subsets based on stratified sampling: nine subsets are used for training, and the remaining samples are used for testing. This evaluation process is repeated 10 times, and the average result is presented. The 10-fold CV results are summarized in Table 9.

Table 9 shows that as the training sample size increases, the performance of these five classification methods is significantly improved. Model based methods LDA+SVM and ICA+SVM perform very well, with the classification accuracy increased significantly. In particular, the prediction accuracy of ICA+SVM ranges from 86.5% to 96.57% in all tumor expression profile datasets, which is comparable with those of SRC, MSRC and PFMSRC.

We can conclude that model-based approaches are more vulnerable to the small sample size problem, over-fitting should be resolved properly.

Discussion

Based on the above experiments, we can draw the following observations:

1. Sparse representation based methods (SRC, MSRC, PFMSRC) consistently outperform the model-based methods (LDA+SVM, ICA+SVM) on all experiments. Especially, in balance splited datasets the prediction accuracy of model-based methods is significantly lower than that of sparse representation methods which may be attributed to the small sample size problem. However, SRC, MSRC, and PFMSRC perform well even when we take 5 samples per subclass for training and the rest for testing.
2. SRC, MSRC and PFMSRC are robust to various sample sizes and feature dimensions, as well as converge fast to the optimal classification rate. The experiments verify the results in [7], which favors the application of those methods. Note that, model-based methods (LDA+SVM, ICA+SVM) exhibit improved 10-fold CV classification accuracy. A reasonable explanation is that the over-fitting phenomena are dramatically reduced when 90% of original samples are used for training and the remaining 10% are used for evaluation in our experiments.
3. PFMSRC outperforms SRC and MSRC in most cases, which implies that the parameter free sparse representation and weighting strategies can capture more discriminating information, especially in multiclass classification. See Figure 5.
4. PFMSRC is a parameter-free method, in which the data dependent sparsity can be self-adaptively tuned, compared with SRC and MSRC in which search for a regularization parameter is laborious work. Moreover, the number of meta-samples is a key parameter for MSRC, as shown in Figure 2, which makes model selection more difficult.

Conclusions

In this study, we proposed a novel non-parametric meta-sample-based sparse representation. The algorithm assumes that test samples can be well reconstructed within a linear combination of weighed meta-samples in the same class. We theoretically proved the rationality of the weighting strategy. A simple but efficient projection function is constructed by the sparse representation coefficients to complete the classification work. We also compare the performance of PFMSRC with that of two model-based methods and two sparse representation-based methods on eight tumor expression datasets. Experimental results have shown the superiority of the proposed method. We then drew some conclusions on the effects of both balanced split and imbalanced split testing/training sets on tumor classification problems.

PFMSRC exhibits stable performance with respect to different training sample sizes and feature dimensions compared with the other four algorithms. Thus, the extension of the sparse representation with dimensionality reduction (feature selection or feature extraction) in a unified framework is one of our future works.

Author Contributions

Conceived and designed the experiments: YJ BL. Performed the experiments: GY ZC. Analyzed the data: LC WZ. Wrote the paper: YJ BL.

References

1. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, et al. (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 403: 503–511.
2. West M (2003) Bayesian factor regression models in the large p, small n paradigm. Bayesian statistics 7: 723–732.
3. Liu B, Fang B, Liu X, Chen J, Huang Z (2013) Large margin subspace learning for feature selection. Pattern Recognition.
4. Cai D, He X, Zhou K, Han J, Bao H (2007) Locality sensitive discriminant analysis. In: IJCAI. pp. 708–713.
5. Sugiyama M (2006) Local fisher discriminant analysis for supervised dimensionality reduction. In: Proceedings of the 23rd international conference on Machine learning. ACM, pp. 905–912.
6. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, et al. (2012) A survey on filter tech-niques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 9: 1106–1119.
7. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31: 210–227.
8. Zheng CH, Zhang L, Ng TY, Shiu CK, Huang DS (2011) Metasample-based sparse representation for tumor classification. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 8: 1273–1282.
9. West M, Blanchette C, Dressman H, Huang E, Ishida S, et al. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. Proceedings of the National Academy of Sciences 98: 11462–11467.
10. Zheng CH, Ng TY, Zhang L, Shiu CK,Wang HQ (2011) Tumor classification based on non-negative matrix factorization using gene expression data. NanoBioscience, IEEE Transactions on 10: 86–93.
11. Zheng CH, Zhang L, Ng V, Shiu CK, Huang DS (2011) Molecular pattern discovery based on penalized matrix decomposition. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 8: 1592–1603.
12. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21: 631–643.
13. Wright GW, Simon RM (2003) A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics 19: 2448–2455.
14. Robnik-Šikonja M, Kononenko I (2003) Theoretical and empirical analysis of relieff and rrelieff. Machine learning 53: 23–69.
15. Seung D, Lee L (2001) Algorithms for non-negative matrix factorization. Advances in neural information processing systems 13: 556–562.
16. Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences 97: 10101–10106.
17. Han X (2010) Nonnegative principal component analysis for cancer molecular pattern discovery. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 7: 537–549.
18. Zheng CH, Huang DS, Zhang L, Kong XZ (2009) Tumor clustering using nonnegative matrix factorization with gene selection. Information Technology in Biomedicine, IEEE Transactions on 13: 599–607.
19. Chen S, Donoho D (1994) Basis pursuit. In: Signals, Systems and Computers, 1994. 1994 Confer-ence Record of the Twenty-Eighth Asilomar Conference on. IEEE, volume 1, pp. 41–44.
20. Donoho DL (2006) Compressed sensing. Information Theory, IEEE Transactions on 52: 1289–1306.
21. Sharon Y, Wright J, Ma Y (2007) Computation and relaxation of conditions for equivalence between l1 and l0 minimization. submitted to IEEE Transactions on Information Theory 5.
22. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological): 267–288.
23. Chen SS, Donoho DL, Saunders MA (1998) Atomic decomposition by basis pursuit. SIAM journal on scientific computing 20: 33–61.
24. Cheng B, Yang J, Yan S, Fu Y, Huang TS (2010) Learning with l1-graph for image analysis. Trans Img Proc 19: 858–866.
25. Grant M, Boyd S, Ye Y (2008). Cvx: Matlab software for disciplined convex programming.
26. Chang CC, Lin CJ (2011) Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2: 27.
27. Vapnik VN (1999) An overview of statistical learning theory. Neural Networks, IEEE Transactions on 10: 988–999.
28. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, et al. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences 96: 6745–6750.
29. Kent ridge bio-medical dataset. Available: http://datam.i2r.a-star.edu.sg/datasets/krbd/. Accessed: 2014 Feb 1.
30. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science 286: 531–537.
31. Gems database. Available: http://www.gems-system.org/. Accessed: 2014 Feb 1.

32. Nutt CL, Mani D, Betensky RA, Tamayo P, Cairncross JG, et al. (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer research 63: 1602–1607.

33. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, et al. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine 7: 673–679.

34. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, et al. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer cell 1: 133–143.

35. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, et al. (2001) Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature genetics 30: 41–47.

36. Wang SL, Zhu YH, Jia W, Huang DS (2012) Robust classification method of tumor subtype by using correlation filters. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 9: 580–591.