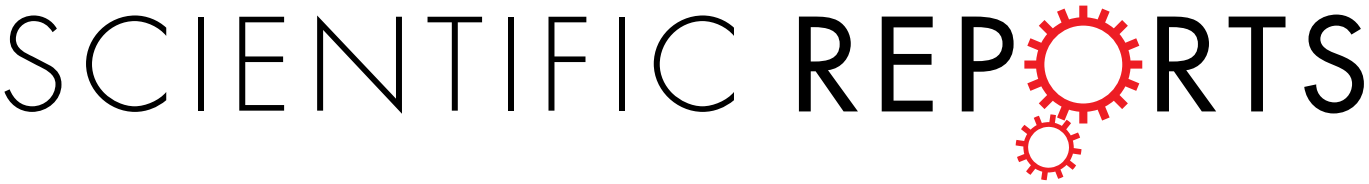

OPEN

Maxdenominator Reweighted Sparse Representation for Tumor Classification

Received: 14 October 2016
Accepted: 08 March 2017
Published: 10 April 2017

Weibiao Li[1], Bo Liao[1], Wen Zhu[1], Min Chen[1], Li Peng[1], Xiaohui Wei[1], Changlong Gu[1] & Keqin Li[2]

The classification of tumors is crucial for the proper treatment of cancer. Sparse representation-based classifier (SRC) exhibits good classification performance and has been successfully used to classify tumors using gene expression profile data. In this study, we propose a three-step maxdenominator reweighted sparse representation classification (MRSRC) method to classify tumors. First, we extract a set of metagenes from the training samples. These metagenes can capture the structures inherent to the data and are more effective for classification than the original gene expression data. Second, we use a reweighted ℓ_1 regularization method to obtain the sparse representation coefficients. Reweighted ℓ_1 regularization can enhance sparsity and obtain better sparse representation coefficients. Third, we classify the data by utilizing a maxdenominator residual error function. Maxdenominator strategy can reduce the residual error and improve the accuracy of the final classification. Extensive experiments using publicly available gene expression profile data sets show that the performance of MRSRC is comparable with or better than many existing representative methods.

Accurate tumor classification is beneficial for cancer treatment. A tumor can be classified as benign, premalignant, or malignant, of which only a malignant tumor can be called cancer. Thus, a reliable and precise classification of tumors is valuable for accurate diagnosis.

Traditional histopathological approach classifies tumors by microscopic tissue examination. However, this process fails to classify many cancer cases. Many classification methods using gene expression profile data have been adapted to classify tumors. Golub *et al.*[1] successfully distinguished acute myeloid leukemia (AML) from acute lymphocytic leukemia (ALL) using gene expression profile data. Furey *et al.*[2] classified cancer tissue samples by support vector machines (SVMs) with gene expression profile data. Bhattacharjee *et al.*[3] precisely classified human lung humors. Ship *et al.*[4] predicted lymphoma by gene expression profile data and supervised machine learning. Huang *et al.*[5] used independent component analysis (ICA)-based penalized discrimination method to classify tumors. Ghosh *et al.*[6] used least absolute shrinkage and selection operator method to classify and select biomarkers in genomic data. All these methods will suffer from overfitting when used to classify tumors.

Sparse representation has been attracting much attention because of the progress in ℓ_1 norm minimization-based methods, such as basis pursuit[7] and compressive sensing theory[8–10]. Wright *et al.*[11] proposed a sparse representation-based classification (SRC) method for face recognition. Common two-stage machine learning methods initially create a training model for testing, and then the samples are tested. By contrast, SRC uses a sparse linear combination of the training samples to represent the testing sample. In this model, a ℓ_1 norm least square method[12] is applied to search for the sparse representation coefficient which will decide the type of the test sample. This essential characteristic would prevent the SRC to perform the training and testing steps, reducing the overfitting problem. Hang *et al.*[13] used SRC to classify tumors and obtained very good experimental results.

However, the original training samples of gene expression profile data may be not efficient to represent the input testing samples. Brunet *et al.*[14] demonstrated that metagenes can recover meaningful biologic information from gene expression profile data. A metagene which can cover the structure inherent to the data is a linear combination of the gene expression profiles of samples. Metagenes can be obtained using singular value decomposition (SVD), principal component analysis (PCA), and nonnegative matrix factorization (NMF). Brunet *et al.*[14] discovered the metagenes by NMF and used these metagenes to cluster the samples, yielding very good

[1]College of Information Science and Engineering, Hunan University, Changsha, Hunan 410082, China. [2]Department of Computer Science, State University of New York, New Paltz, New York 12561, USA. Correspondence and requests for materials should be addressed to B.L. (email: dragonbw@163.com)

results. Zhen *et al.*[15] used SVD or NMF to capture the metagenes from the gene expression profile data. In the current study, we determined the metagenes using SVD. Several studies have reported on the extraction of the metagenes. NMF was used in ref. 14 to decompose gene expression patterns as an additive combination of a few metagene patterns. The NMF metagenes could overlap and thus expose the participation of a single gene in multiple pathways or processes. ICA was applied in ref. 16 to gene expression data, and a linear model, which was termed "expression modes," was derived. In this model, the expression of each gene is a linear function of the expression modes. The dominant expression modes were experimentally found to be related to distinct biological function, such as the phase of the cell cycle or the mating response. SVD was applied in ref. 17 to transform genome-wide expression data from "genes" × "arrays" space to reduced diagonalized "eigengens" × "eigenarrays" space, where the eigengenes (or metagenes) are unique orthonormal superposition of the genes (or samples). SVD was applied to capture the weighted metagenes[18], then the test sample is represented as the linear combination of these weighted metagenes. These analytical results show that using metagenes to replace gene expression data can produce better results for classification.

In addition, ℓ_1 norm is just as a replacement of the ℓ_0 norm in the objective function of sparse representation. However, ℓ_1 and ℓ_0 norms significantly differ. According to the definitions of ℓ_1 and ℓ_0 norms, larger coefficients are penalized more heavily in the ℓ_1 norm than smaller coefficients, contrary to the more democratic penalization of the ℓ_0 norm. Candès *et al.*[19] addressed this problem by proposing a reweighted ℓ_1 norm minimization method for signal recovery. In the current study, we also use the reweighted ℓ_1 norm minimization strategy to find the sparse representation coefficients.

We propose a maxdenominator residual error function which takes full advantage of the linear relation between test sample and metagenes to capture the classification from the sparse representation coefficients.

We first extract a set of metagenes from the training samples. Each testing sample is represented as a linear combination of metagenes, and the linear combination coefficients of metagenes are captured by reweighted ℓ_1 norm minimization strategy. Then, we use a maxdenominator residual error function to obtain the classification result. Our method is named as Maxdenominator Reweighted Sparse Representation Classification (MRSRC).

The rest of this paper is organized as follows. In the Result Section, the experiments will be presented, and the performance of the proposed method will be compared with those of several common methods for tumor classification. In the Discuss Section, the results are discussed, and conclusions are drawn. Finally, we will describe the metagene model and briefly review SRC. We will also introduce the reweighted sparse representation method and the maxdenominator residual error function. We will also describe the details of MRSRC, including its object function and the solution algorithm in the Method Section.

Results

In this section, we run an experiment to evaluate the performance of the proposed MRSRC method. The experiment is divided into five parts. The first part involves the two-class classification given in the Two-Class Classification Section. The second part is the multi-class classification given in the Multiclass Classification Section. The third part comprises the different feature dimensions given in the Experiment with different number of genes Section. The fourth part involves the cross-validation given in the Comparison of CV performance Section. The fifth part visually presents the sparse representation coefficients given in the Visualization of Sparse Representation Coefficients Section. The proposed method is compared with several state-of-the art methods, such as the widely used LDA + SVM[20], ICA + SVM[21], SRC[13], and SVD + MRSC[15], MACE[22], OTSDF[22]. The former two methods are models based on SVM as a classification and accompanied by feature extraction. We use these two methods as the baseline. The latter two methods are template based and get very good performance for classification of tumors.

In our proposed MRSRC method, three parameters should be set, namely, positive regularization parameter λ, stability parameter ε, and maximum reweighted iterations parameter$\ell_{\max}$. According to a previous study[23], the value of λ is given as follows:

$$\lambda = 0.01\lambda_{\max}, \tag{1}$$

$$\lambda_{\max} = \|2X^T x\|_\infty, \tag{2}$$

where $\|\mu\|_\infty = \max |\mu_i|$ denotes the ℓ_∞ norm of the vector μ. For the stability parameter ε, as a rule of thumb, ε should be slightly smaller than the expected nonzero magnitudes of θ. From Fig. 1, we can set $\varepsilon = 0.1$ in this paper. Finally, we set the maximum reweighted iterations parameter $\ell_{\max} = 4$. These values were selected because much of the benefit comes from the first few reweighting iterations, and the added computational cost for improved sparse representation coefficients is quite moderate.

The SVM kernel parameters for the LDA + SVM and ICA + SVM algorithms are determined by 10-fold cross validation. In addition, we simply extract $c-1$ (where c denotes the number of classes) new features to train the classifier, because LDA can find a maximum of $c-1$ meaningful projection vectors in the subspace. Moreover, the determination of the number of independent components of ICA is also an empirically dependent work. Here, we use the same method as suggested by ref. 24 Zheng *et al.* The SRC and MSRC methods also need parameter λ to control sparsity. The parameters of methods of MACE and OTSDF are set as recommended by[22] Wang *et al.*

To avoid the effects of imbalanced training set, we use Balance Division Method (BDM) to divide each original data set into balanced training set and test set. For this BDM, Q samples from each subclass are randomly selected to be used for training set, and the rest are used for test set. Here, Q must be an integer number. For example, if we set Q to 5, then 5 samples per subclass randomly selected are used as training set and the rest are assigned to the test set. Then we evaluate the performance of seven methods on a balanced split data set. We randomly select

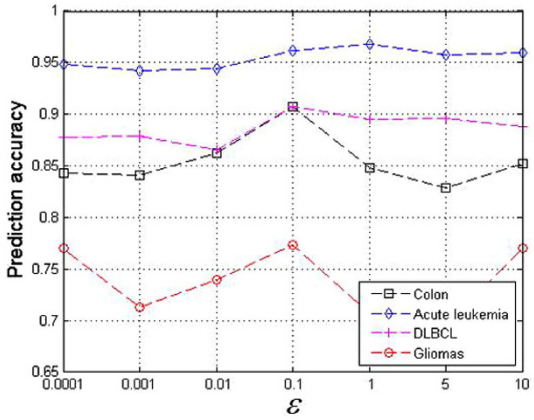

Figure 1. Optimal classification accuracy of MRSRC on four binary class dataset.

Data set	Classes	Genes	The number of samples
Acute leukemia data	2	7,129	72
Colon cancer data	2	2,000	62
Gliomas data	2	1,2625	50
DLBCL data	2	7,129	77

Table 1. The descriptions of four data sets for two-class classification.

$p = 5$ 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. In the test, p denotes the number of training samples per classes, and $\min(|c_i|)$ denotes the minimum number of subclass set of samples in the training data. The training/testing are performed 10 times, and the average classification accuracies are presented.

To evaluate the classification performance in imbalanced split training/testing sets, we perform a 10-fold stratified CV on a tumor subtype data sets. All samples are randomly divided into 10 subsets on the basis of stratified sampling: nine subsets are used for training, and the remaining samples are used for testing.

We determine the accuracy, sensitivity, and specificity to measure the performance for a fair experimental evaluation. These metrics are defined as follows:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}, \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{4}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{5}$$

where true positives (TP) and true negatives (TN) denote the ratio of samples that are correctly classified into the positive and negative groups, respectively. False negatives (FN), and false positives (FP) denote the incorrectly classified true positive samples into the negative group and true negative samples into the positive group, respectively.

Two-Class Classification. In this section, four publicly available microarray data sets are used to compare the performance of the methods. These data sets are acute leukemia data set[1], colon cancer data set[25], gliomas data set[26], and diffuse large B-cell lymphoma (DLBCL) data set[27] (please see supplementary information for data sets).

The acute leukemia data set contains 72 samples of 7,129 genes. The target class has 2 states, including 47 acute lymphoblastic leukemia patients and 25 acute myelogenous leukemia patients. The colon cancer data set includes 40 tumor and 22 normal colon tissue samples and contains the expression of 2,000 genes with the highest minimal intensity across 62 tissues. The gliomas data set consists of 50 samples with two subclasses (glioblastomas and anaplastic oligodendrogliomas), and each sample contains 12,625 genes. The DLBCL data set contains 77 samples of 7,129 genes. The target class has 2 states, including 58 diffuse large b-cell lymphoma samples and 19 follicular lymphoma samples. Table 1 provides the details of the data sets.

The average prediction accuracies of the classification of the balanced training set and test set are shown in Fig. 2. The MRSRC exhibits encouraging performance. The MRSRC achieves the best classification accuracy in all cases for the colon cancer data (Fig. 2b). Although gliomas are difficult to classify, the proposed approach can still

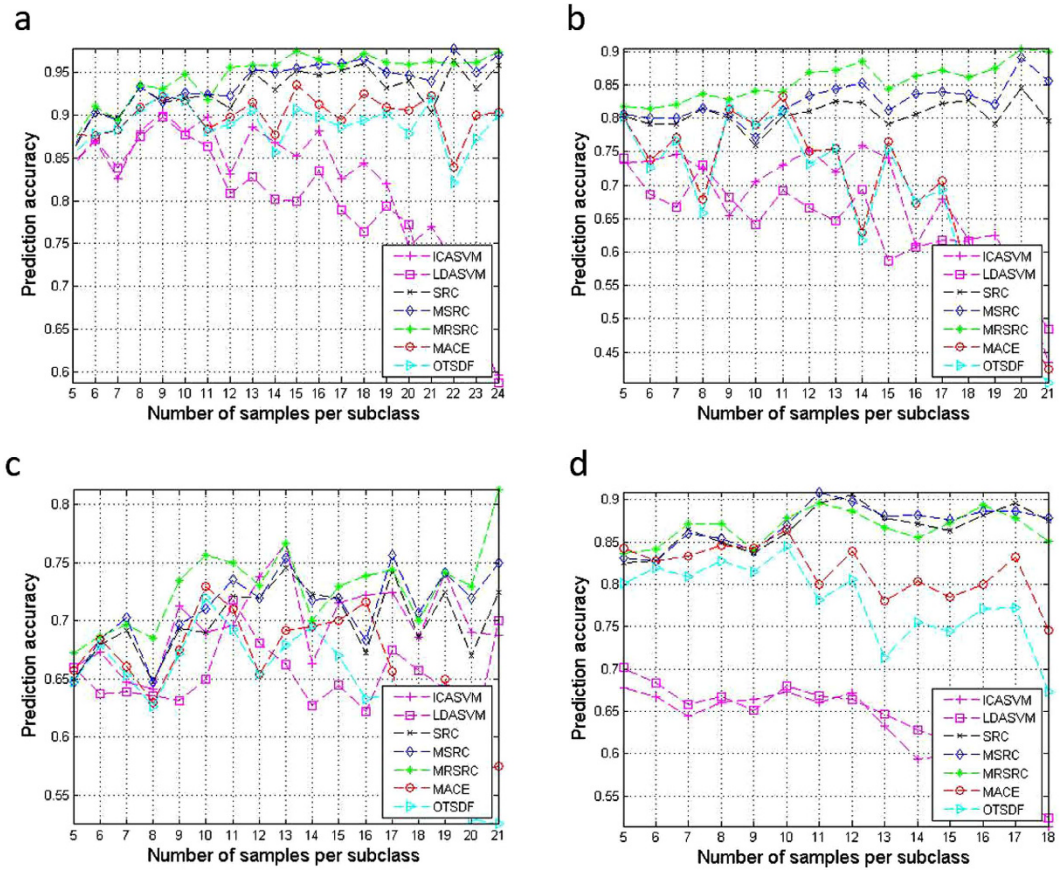

Figure 2. Comparison of prediction accuracy on four binary classification datasets by varying the number of samples from per subclass.

Data set	SRC	MSRC	MRSRC
Acute leukemia data	**94.00%**	93.33%	**94.00%**
Colon cancer data	80.00%	85.83%	**87.50%**
Gliomas data	67.22%	**67.78%**	**67.78%**
DLBCL data	**96.67%**	93.33%	93.33%

Table 2. The classification sensitivity of two-class classification when the numbers of metagenes per subclass are fixed as 10.

Data set	SRC	MSRC	MRSRC
Acute leukemia data	93.24%	94.59%	**95.41%**
Colon cancer data	77.33%	84.67%	**85.33%**
Gliomas data	70.00%	**73.33%**	72.50%
DLBCL data	88.54%	**89.38%**	88.96%

Table 3. The classification specificity of two-class classification when the numbers of metagenes per subclass are fixed as 10.

achieve the highest classification accuracy via 21 (81%) samples per subclass used for training (Fig. 2c). MRSRC also achieves relatively high prediction accuracies in most cases in the acute leukemia and DLBCL data sets (Fig. 2a and d). Notably, the classification accuracies of LDA + SVM and ICA + SVM drop quickly as the number of samples considered for training increases. These results are consistent with the observations in the literature[22].

We illustrate the experimental results by showing the sensitivity and specificity of the proposed MRSCR, SCR, and MSCR in Tables 2 and 3 when the numbers of metagenes per subclass are fixed to 10.

The experimental results and analysis show that the proposed MRSRC is highly competitive for two-class tumor classification.

Data set	Classes	Genes	samples
SRBCT data	4	2,308	83
ALL data	6	12,625	248
MLLLeukemia data	3	12,582	72
LukemiaGloub data	3	7,129	72

Table 4. The descriptions of four data sets for multiclass classification.

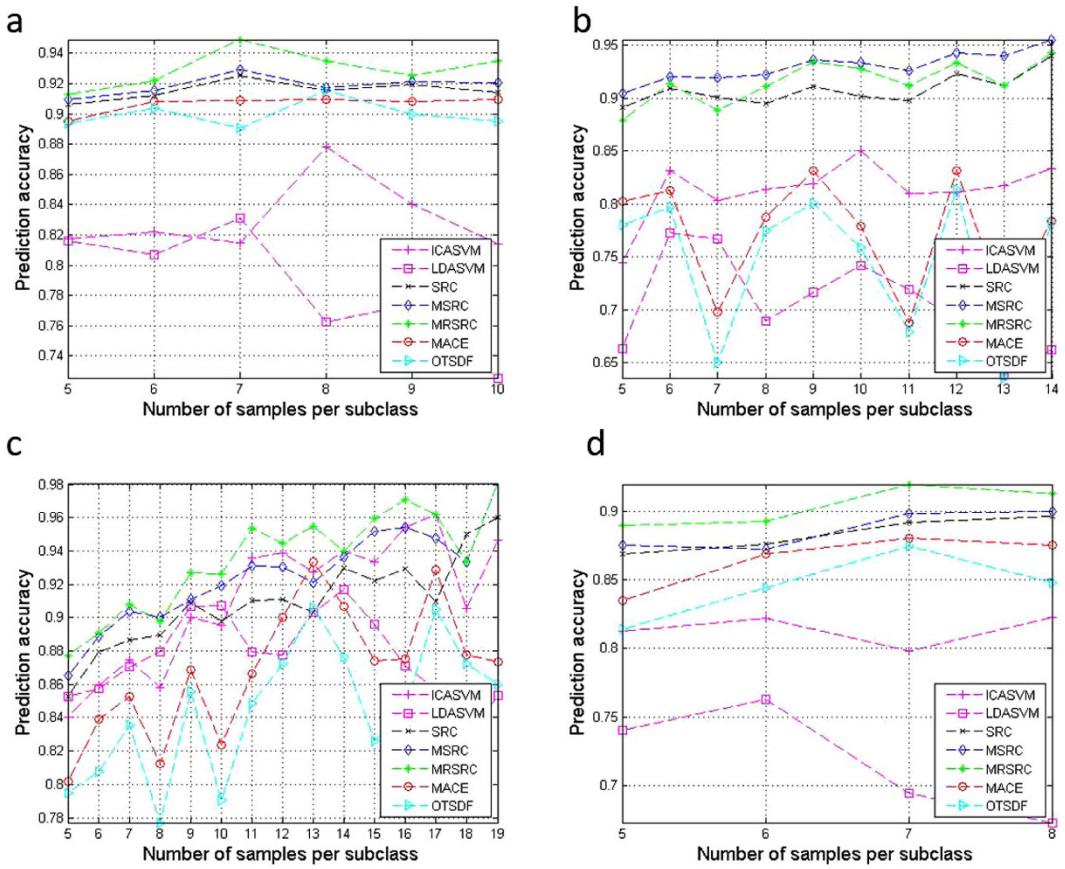

Figure 3. Comparison of prediction accuracy on four multiclass classification datasets by varying the number of samples from per subclass.

Multiclass Classification. In this section, four multiclass data sets are used to further investigate the performance of the MRSRC. The experimental setup is the same as that for the binary classification case. The data sets include the small, round blue cell tumors (SRBCT)[28], ALL[29], MLLLeukemia[30], and LukemiaGloub[1] (please see supplementary information for data sets).

The SRBCT data set is composed of four types of tumors. This data set includes 83 samples with 2,308 genes. The ALL data set consists of 248 samples with six subclasses. Each sample contains 12,625 genes. The MLLLeukemia data set consists of 72 samples with three subclasses. Each sample contains 12,582 genes. The LukemiaGloub data set consists of 72 samples with three subclasses. Each sample contains 7,129 genes. Table 4 provides details of the data sets.

The average prediction accuracy of the classification is shown in Fig. 3. The MRSRC achieves the best classification accuracy in all cases in the SRBCT data and LukemiaGloub data (Fig. 3a and d). In the ALL experiments (Fig. 3b), the proposed MRSRC shows no evident advantages over SRC and MSRC. The MRSRC achieves extremely high prediction accuracy in most cases for the MLLLeukemia data set (Fig. 3c). As shown by the data dimensions in the four experiments, the MRSCR offers more advantages and more stability than other methods.

Experiment with different numbers of genes. In this subsection, we evaluate the performance of the seven 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 the test. We perform the test with various numbers of genes, starting from 102 to 352 genes in increments of 5. The top genes are selected from each data set by applying the Relief-F algorithm[31] to the training set.

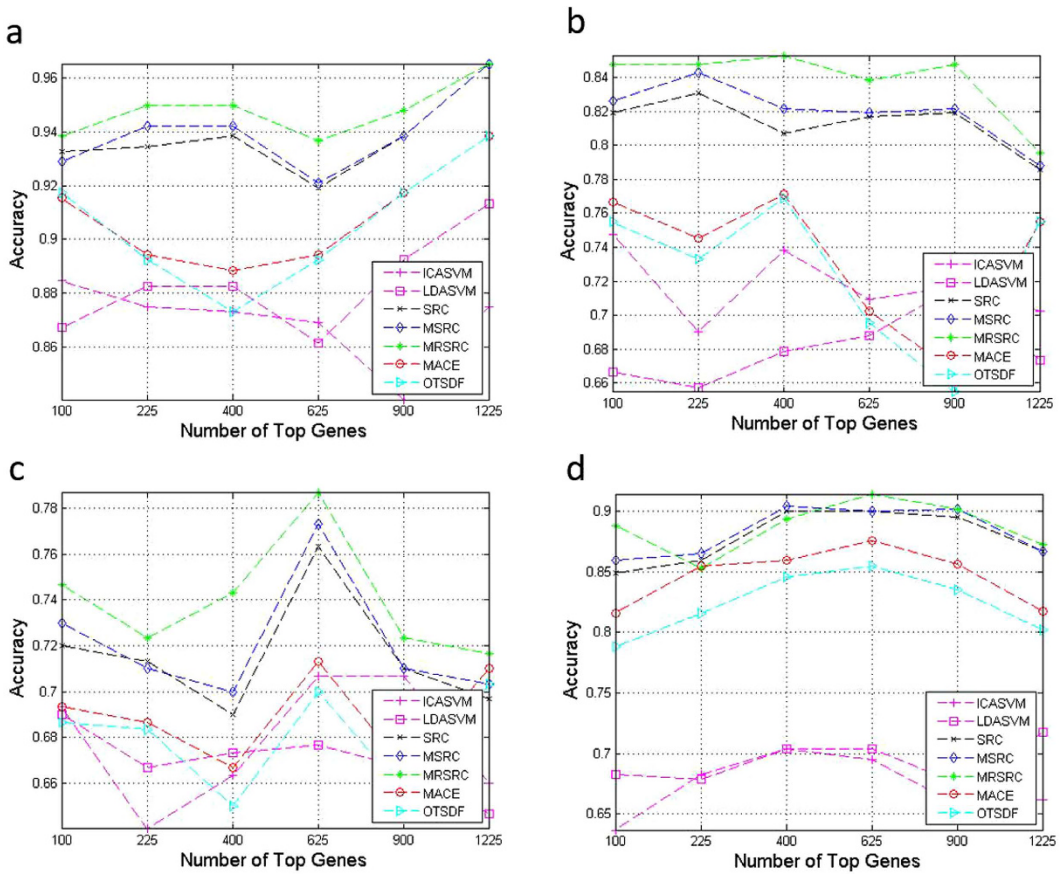

Figure 4. Comparison of accuracy on four binary classification datasets by varying the number of top selected genes.

Figure 4 presents the result of the binary classification. The MRSRC outperforms the other methods in terms of prediction accuracy for all data sets, except for the DLBCL data set (Fig. 4d). The gene selection of the MRSRC is better than that of other methods for the acute leukemia, colon, and gliomas data sets (Fig. 4a,b and c). The MRSRC, SRC, and MSRC share the same curve trend. Evidently, the MRSRC, SRC, and MSRC consistently outperform LDASVM, ICASVM, MACE, and OTSDF in all data sets.

Figure 5 shows the average prediction accuracy of multiclass classification. As shown in Fig. 5, the MRSRC is robust with respect to the number of top-ranked genes. Selection can obviously improve the classification accuracies of all methods in the four data sets. The results show remarkable progress, especially for LDASVM. This result is due to the fact that LDA can capture a large amount of discriminating information in the multiclass classification task and significantly reduce the over-fitting phenomenon in comparison with its performance in the binary classification task.

Comparison of CV performance. To evaluate the classification performance of the MRSRC, SRC, and MSRC in the imbalanced split training/testing sets, we perform a 10-fold stratified CV on the tumor subtype data set. All samples are randomly divided into 10 subsets on the basis of 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 Tables 5, 6 and 7.

Table 5 Shows that the MRSRC achieves the best accuracy in the seven data sets. Especially in the case of multiclass classification, the MRSRC shows superior performance in all data sets. Table 6 shows that the MRSRC achieves the highest value in all eight data sets. Table 7 shows that the MRSRC achieves the highest specificity in all the multiclass.

We can conclude that the MRSRC is outstanding in the imbalance split training/testing sets.

Visualization of Sparse Representation Coefficients. To further analyze the results, we compared the value of coefficients of MSRC and MRSRC for the eight data sets. Figures 6 and 7 show the value of the sparse representation coefficients of the training samples. A test samples is represented as the linear relation of metagenes of such training samples. From Figs 6 and 7, we can observe that MRSRC obtains better sparse representation coefficients. MRSRC demonstrates more coefficients, which are equal to zero and are close to the theoretical results.

Figures 6 and 7 Also show that the test sample can be represented as the linear combination of metagenes. The figures illustrate that metagenes can recover meaningful biological information from gene expression profile data

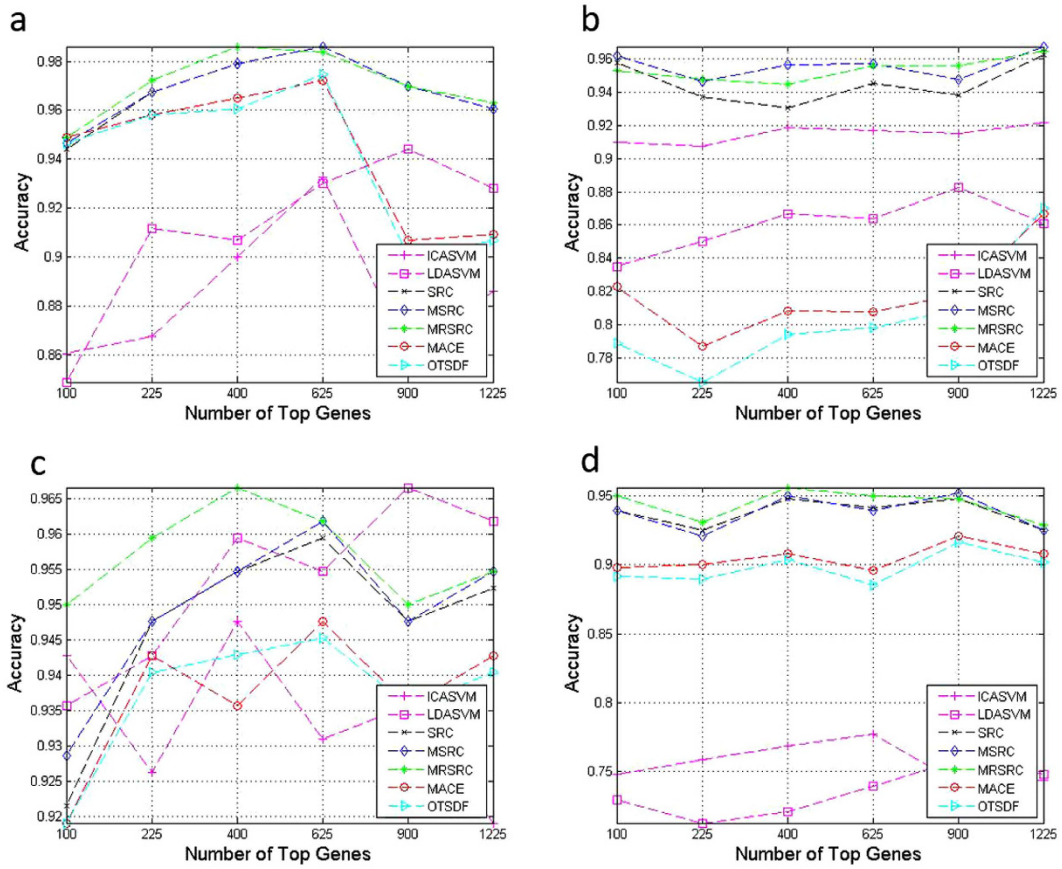

Figure 5. Comparison of accuracy on four multiclass classification datasets by varying the number of top selected genes.

Data set	SRC	MSRC	MRSRC
Acute leukemia data	**83.87%**	**83.87%**	**83.87%**
Colon cancer data	95.83%	97.22%	**98.61%**
Gliomas data	72.00%	72.00%	**78.00%**
DLBCL data	**98.70%**	96.10%	92.21%
SRBCT data	**98.80%**	96.39%	**98.80%**
ALL data	97.98%	97.58%	**98.39%**
MLLLeukemia data	**98.61%**	**98.61%**	**98.61%**
LukemiaGloub data	95.83%	**97.22%**	**97.22%**

Table 5. 10-fold CV prediction accuracy of eight tumor microarray datasets using different classification methods.

Data set	SRC	MSRC	MRSRC
Acute leukemia data	72.73%	**77.27%**	**77.27%**
Colon cancer data	92.00%	92.00%	**96.00%**
Gliomas data	71.43%	71.43%	**78.57%**
DLBCL data	94.74%	94.74%	**100.0%**
SRBCT data	**100.0%**	96.55%	**100.0%**
ALL data	86.67%	86.67%	**93.33%**
MLLLeukemia data	**100.0%**	**100.0%**	**100.0%**
LukemiaGloub data	88.89%	88.89%	**100.0%**

Table 6. 10-fold CV prediction sensitivity of eight tumor microarray datasets using different classification methods.

Data set	SRC	MSRC	MRSRC
Acute leukemia data	**90.00%**	87.50%	87.50%
Colon cancer data	97.87%	**100.0%**	**100.0%**
Gliomas data	72.73%	72.73%	**77.27%**
DLBCL data	**100%**	96.55%	89.66%
SRBCT data	**100%**	**100%**	**100%**
ALL data	98.71%	**99.14%**	**99.14%**
MLLLeukemia data	**100%**	**100%**	**100%**
LukemiaGloub data	**100%**	**100%**	**100%**

Table 7. 10-fold CV prediction specificity of eight tumor microarray datasets using different classification methods.

and can cover the structure inherent to the data. Each metagene in the data set obtained maximum coefficient, which indicates that the test sample is very similar to the metagene.

Discussions

Sparse representation-based methods (SRC, MSRC, and MRSRC) consistently outperform the model-based methods (LDASVM and ICASVM) and the template-based methods (MACE and OTSDF) in most experiments. This result is probably due to the small sample size problem for model-based strategies. Sparse representation-based methods perform well even when five samples per subclass were considered for training and the rest for testing. Moreover, MRSRC outperforms SRC and MSRC in most cases, which implies that reweighted and maxdenominator strategies can contribute to improve sparse solution and obtain better classification results. Additionally, gene selection can enhance the accuracy of classification using all datasets. Gene expression profiling involves data with high dimensionality, and exclusion of the redundant is critical for classification.

In this paper, we propose a new sparse representation-based method for classifying tumors. This method adapts reweighted strategy to balance $\ell_1 - norm$. The reweighted strategy contributes to capture sparse solution, which is used for classification. Classification is achieved by a maxdenominator residual error function, which takes full advantage of the linear relations between the testing sample and metagenes extracted using SVD from the training samples. We also compare the performance of MRSRC with those of two sparse representation-based methods, two model-based methods and two template-based methods using eight tumor expression datasets. The results have shown the superiority of MRSRC and validated the effectiveness and efficiency of MRSRC in tumor classification.

MRSRC exhibits stable performance with respect to different training sample sizes compared with the other four methods. The properties of this reweighted sparse representation algorithm should be investigated further. Thus, we will extend the algorithm with dimensionality reduction in our future studies.

Methods

Metagenes of Gene Expression File Data. Metagenes of gene expression data are captured by mathematical operation on the original sample data. These mathematical operations include SVD, PCA, NMF, ICA, or other linear or nonlinear models. Essentially, mathematical operation on gene expression data contributes to highlight particular biological functions, recover meaningful biological information, capture alternative structures inherent to the data, provide biological insight, reduce noise, and compress the data in a biologically sensible way[14–18]. Another viewpoint presented in ref. 15 shows that the gene expression pattern can be approximately represented as linear combination of these metagenes.

Assuming that $X \in R^{m \times n}$ is a training sample set of gene expression file data, then matrix X has the following approximation by mathematical operation:

$$X = \Phi H, \tag{6}$$

where Φ is a metagene matrix with size of $m \times p$, and each of the p columns are regarded as a metagene. Matrix H is a pattern matrix with size of $p \times n$, and each of the n columns are treated as the metagene expression pattern of the corresponding sample. We will use Φ, instead of X, for the classification.

Several studies have reported on the extraction of the metagenes. NMF was used in ref. 14 to decompose gene expression patterns as an additive combination of a few metagene patterns. The NMF metagenes could overlap and thus expose the participation of a single gene in multiple pathways or processes. ICA was applied in ref. 16 to gene expression data, and a linear model, which was termed "expression modes," was derived. In this model, the expression of each gene is a linear function of the expression modes. The dominant expression modes were experimentally found to be related to distinct biological function, such as the phase of the cell cycle or the mating response. SVD was applied in ref. 17 to transform genome-wide expression data from "genes" × "arrays" space to reduced diagonalized "eigengens" × "eigenarrays" space, where the eigengenes (or metagenes) are unique orthonormal superposition of the genes (or samples). SVD was applied to capture the weighted metagenes[18], then the test sample is represented as the linear combination of these weighted metagenes.

These analytical results show that using metagenes to replace gene expression data can produce better results for classification.

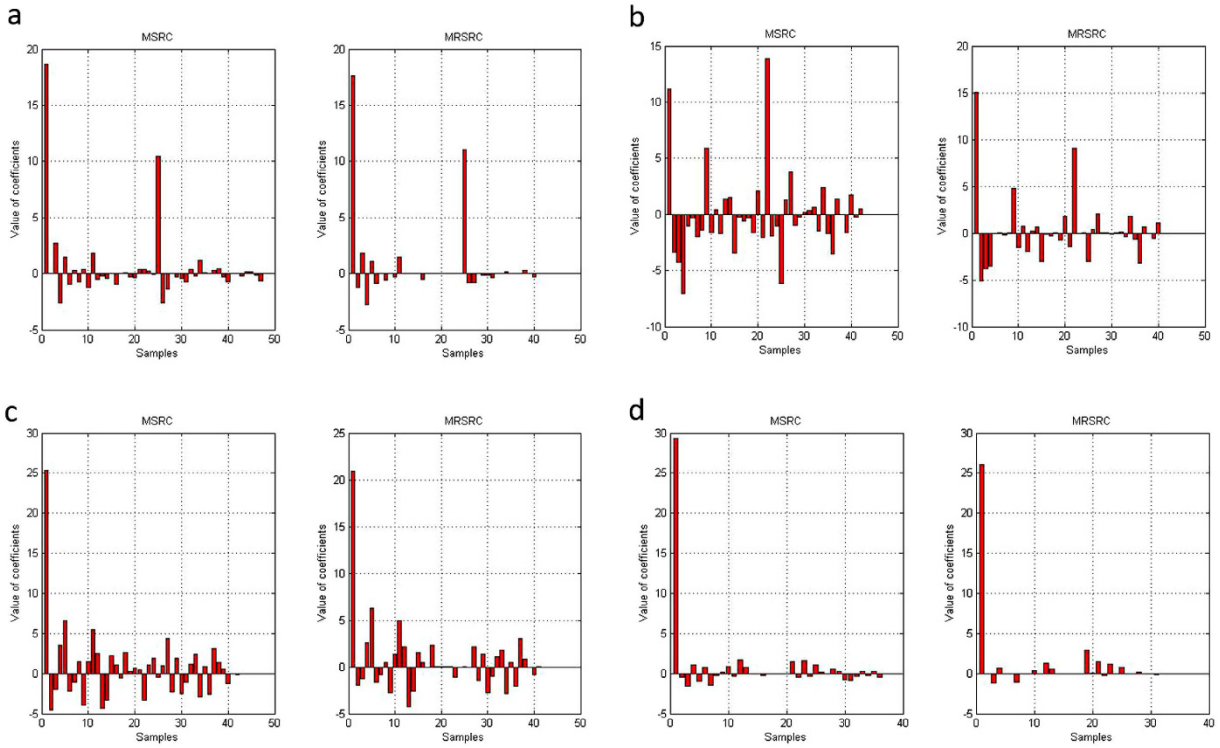

Figure 6. The value of the sparse representation coefficients of MSRC and MRSRC on four binary classification datasets when choosing one sample as test set.

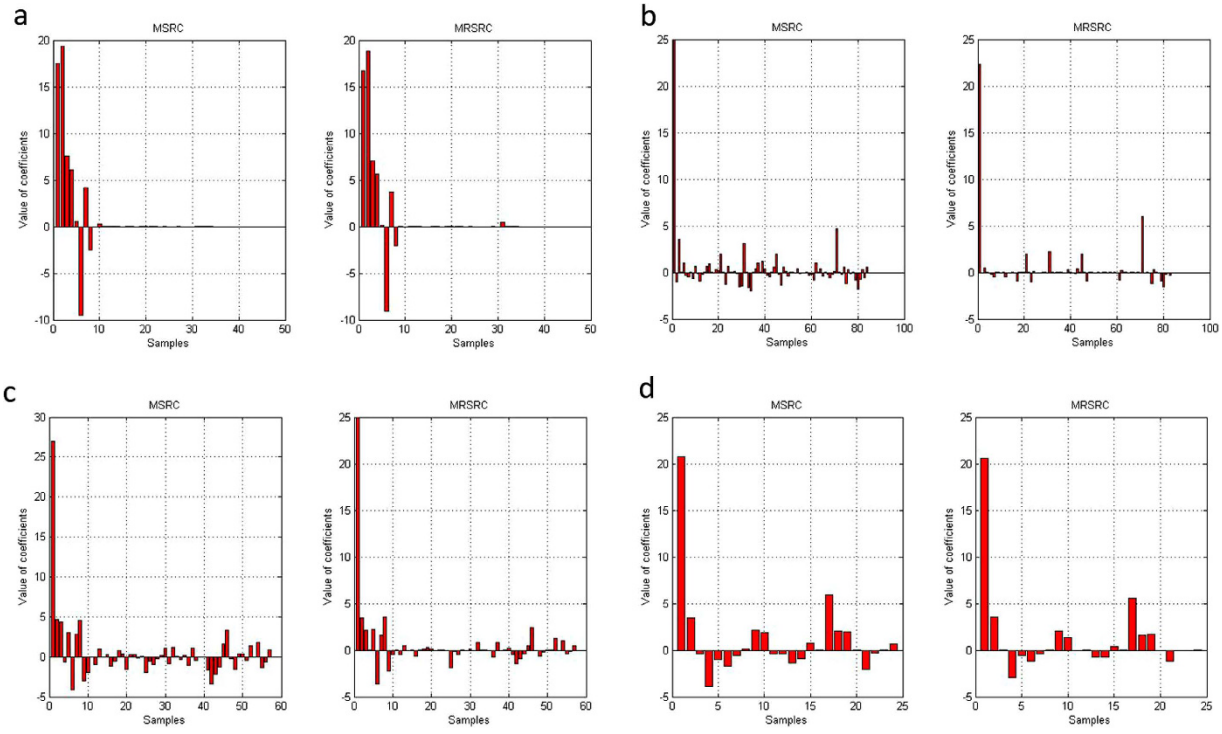

Figure 7. The value of the sparse representation coefficients of MSRC and MRSRC on four multiclass classification datasets when choosing one sample as test set.

Sparse Representation for the Classification of Testing Tumor Samples. SRC has been successfully applied to tumor classification in refs 13 and 15. The experimental results showed that SRC is quite competitive in

classifying tumors. We simply introduce several notations of this model to enhance the understanding of the SRC. We assume that $\{X_i | X_i \subseteq R^m, i = 1, 2, \ldots, c\}$ is a set of training samples, and $\{y_i | y_i \in \{1, 2, \ldots, c\}\}$ is labeled corresponding to X_i, where m is the dimensionality of samples, and c is the number of classes. The *jth* class training samples X_j can be presented as columns of a matrix $X_j = [x_{j,1}, x_{j,2}, \ldots, x_{j,n_j}] \in R^{m \times n_j}$, $j = 1, \ldots, c$, where $x_{j,I}$ is a sample of the *jth* class, and n_j is the number of the *jth* class training samples. Thus, a training sample matrix X can be obtained, as follows:

$$X = [X_1, X_2, \ldots, X_c] \in R^{m \times n}, \tag{7}$$

where $n = \sum_{j=1}^{c} n_j$.

Then, we use SVD to obtain the metagenes of each class. We factorize each class training sample matrix X_i, as follows:

$$X_i = \Phi_i H_i. \tag{8}$$

We can derive the metagenes from the c classes after computing the metagenes Φ_i of each class, as follows:

$$\Phi = [\Phi_1, \ \Phi_2 \ldots, \ \Phi_c]. \tag{9}$$

Given a test sample $x \in R^m$, according to SRC, x can be represented as the linear combination of metagenes of all training samples, as follows:

$$x = \Phi\theta, \tag{10}$$

where $\theta \in R^n$ is the sparse representation coefficients. The x can also be represented in the form of an equation, as follows:

$$x = \Phi_1\theta_1 + \Phi_2\theta_2 + \ldots + \Phi_c\theta_c, \tag{11}$$

where $\theta_i \in R^{n_i}$ is the coefficient vector corresponding to the *ith* class training samples. Ideally, if x belongs to the *jth* class, then θ has the following form:

$$\theta = \left[0, \ldots, 0, \underbrace{\theta_{j,1}, \ldots, \theta_{j,n_j}}_{\theta_j}, 0, \ldots 0\right]^T. \tag{12}$$

Equation (12) indicates that only the entries corresponding to the *jth* training sample X_j are not zeros. Then, x can be finally expressed as follows:

$$x = \theta_{j,1}\Phi_{j,1} + \ldots + \theta_{j,n_j}\Phi_{j,n_j}, \tag{13}$$

where $\Phi_{j,i}$ is a metagene of the *jth* class training samples, and n_j is the number of the *jth* class training samples.

Then, the key problem becomes the calculation of the sparse representation vector θ. Only a fraction of entries are not zeros, so θ is expected to be sparse. Several excellent methods have been put forward in the theory of sparse representation and compressive sensing[8–10].

Reweighted Sparse Representation. The sparse representation coefficients of a sample $x \in R^m$ can be obtained mathematically by solving the combinatorial optimization problem, as follows:

$$\min_{x \in R^m} \|\theta\|_{\ell_0} \ subject \ to \ x = \Phi\theta, \tag{14}$$

where $\|\theta\|_{\ell_0}$ is the ℓ_0 norm. However, the optimization problem given by Equation (14) is NP-hard and generally impossible to solve efficiently, because its solution usually requires an intractable combinatorial search[32].

A common alternative to the combinatorial optimization problem is to solve the following convex problem:

$$\min_{x \in R^m} \|\theta\|_{\ell_1} \ subject \ to \ x = \Phi\theta, \tag{15}$$

where $\|\theta\|_{\ell_1} = \sum_{i=1}^{n} |\theta_i|$, is the ℓ_1 norm. Contrary to Equation (14) and (15) can actually be recast as a linear program and can be solved efficiently[33]. The distinction between Equations (14) and (15) is the choice of objective function, with the latter using ℓ_1 norm as a proxy for the literal ℓ_0 norm sparsity count. However, a key difference exists between the ℓ_1 and ℓ_0 norms. As mentioned earlier, the coefficients are penalized when they are imbalanced in the ℓ_1 norm, contrary to the more equal penalization in the ℓ_0 norm.

First, we consider the following weighted ℓ_1 minimization problem to resolve the important issues:

$$\min_{\theta \in R^n} \sum_{i=1} w_i |\theta_i| \ subject \ to \ x = \Phi\theta, \tag{16}$$

where w_i is the positive weight. Similar to Equation (15) and (16) is convex and can be recast as a linear program. For convenience, Equation (16) can be rewritten as follows:

$$\min_{\theta \in R^n} \|W\theta\|_{\ell_1} \ subject \ to \ x = \Phi\theta, \tag{17}$$

where W is a diagonal matrix with $w_1 \ldots w_m$ on the diagonal and zeros elsewhere.

Equation (17) can be viewed as a relaxation of the following weighted ℓ_0 minimization problem:

$$\min_{x \in R^n} \|W\theta\|_{\ell_0} \ subject \ to \ x = \Phi\theta. \tag{18}$$

If the solution to Equation (14) is unique, then the mathematical solution to Equation (18) is unique, and the weights do not vanish[19]. However, this situation has changed for the corresponding ℓ_1 relaxations Equations (15) and (16). These equations will generally have different solutions. Thus, how to set the weights (w_i) as free parameters in the convex relaxation is important to improve signal reconstruction.

We set the weights (w_i) inversely to the magnitude of θ_i, as follows:

$$w_i = \frac{1}{|\theta_i| + \varepsilon}, \tag{19}$$

where parameter $\varepsilon > 0$ to provide stability and ensure that a zero-valued component in θ_i does not strictly prohibit a nonzero estimate at the next step. Thus, $w_i\theta_i \approx 1$ and $\|W\theta\|_{\ell_1} \approx \|\theta\|_{\ell_0}$. Thus, Equations (14) and (15) present slight difference.

An iterative algorithm that alternates between estimating θ and redefining the weights is proposed in ref. 19. We have an updated algorithm of θ_i and w_i as follows:

1. The initial value of iteration count ℓ is set to zero, and weights are set to $w_i^{(0)} = 1$, $i = 1, \ldots, m$.

2. The weighted ℓ_1 minimization problem is solved as follows:

$$\theta^{(\ell)} = \text{argmin}\|W^{(\ell)}\theta\|_{\ell_1} \ subject \ to \ x = \Phi\theta \tag{20}$$

3. The weights are updated, such that for each $i = 1, \ldots, m$,

$$w_i^{(\ell+1)} = \frac{1}{|\theta_i^{(\ell)}| + \varepsilon}. \tag{21}$$

4. When ℓ attains a specified maximum number of iterations $\ell_{\max}$ or Equation (20) is terminated upon convergence. Otherwise, ℓ is incremented, and step 2 is reiterated.

In step 4, the parameter of $\ell_{\max}$ is the maximum number of reweighted iterations.

Maxdenominator Residual Error Function. The sparse representation coefficient θ can be obtained using the iterative algorithm. Ideally, θ has a form similar to Equation (12) with nonzero entries corresponding to the class of the testing sample. However, modeling error and noise will inevitably lead to small nonzero entries corresponding to multiple object classes[11]. For a more robust classification, x is classified based on how well x can be reconstructed using the coefficients from each class[34]. For class i, we define a characteristic function δ_i, as follows:

$$\delta_i(\theta_j) = \begin{cases} \theta_j, \text{if } y_j = i \\ 0, \text{ otherwise} \end{cases} \tag{22}$$

where y_j is the label of training sample X_j. This function is used to obtain the ideal coefficients of samples belonging to class i and a new vector, as follows:

$$\vartheta_i = [\delta_i(\theta_1), \delta_i(\theta_2), \ldots, \delta_i(\theta_c)]^T. \tag{23}$$

In addition, the test sample is represented as the linear combination of the metagenes in sparse representation. Thus, a metagene from the same class of the test sample is very similar to the test sample. The sparse representation coefficient of this metagene is relatively large. Then, the relatively large coefficient is used to reduce the residual error.

Finally, a maxdenominator residual error function is defined, as follows:

$$r_i(x) = \|\Phi\vartheta_i - x\|_2 / \max(\vartheta_i), \tag{24}$$

where $\max(\vartheta_i)$ yields the max coefficients of ϑ_i. Thus, the label $\hat{y}$ of x can be estimated by minimizing the following formula:

$$\hat{y} = \arg \min_{i=1,\ldots,c} r_i(x) = \|\Phi\vartheta_i - x\|_2 / \max(\vartheta_i) \tag{25}$$

The truncated Newton interior-point method[23] is used to solve the optimization problem in Equation (17). The solution was performed using the $\ell_1 - \ell_s$ Matlab package available online (https://stanford.edu/~boyd/l1_ls).

The complete classification algorithm of MRSRC is as follows:

Input: Training samples $X = [X_1, X_2, \ldots, X_C] \in R^{m \times n}$

Testing samples $x \in R^m$
Output: Label $\hat{y}$ of x
Procedure
1. Columns of X are normalized to have unit ℓ_2-norm.
2. The metagenes of every class are extracted by SVD.
3. The initial value of iteration count ℓ is set to zero, and weights are $w_i^{(0)} = 1,\ i = 1,\ \ldots,\ m$.
4. The weighted ℓ_1 minimization problem Equation (17) is solved to obtain the coefficient vector$\theta^{(\ell)}$.
5. The weights are updated according to Equation (21), that is, for each $i = 1, \ldots, m$,.
6. When ℓ attains a specified maximum number of iterations $\ell_{\max}$ or Equation (20) is terminated upon convergence. Otherwise, ℓ is incremented, and step 2 is reiterated.
7. Count $\vartheta_i = [\delta_i(\theta_1),\ \delta_i(\theta_2),\ \ldots,\ \delta_i(\theta_c),\ ]^T,\ i = 1,\ 2,\ \ldots,\ c$.
8. The c classes residual errors are calculated, that is, $\mathrm{r}_i(x) = \left\|\Phi\vartheta_i - x\right\|_2 / \max(\vartheta_i),\ i = 1,\ 2,\ \ldots,\ c$.
9. The label $\hat{y}$ of x is estimated according to residual errors.

References

1. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537 (1999).
2. Furey, T. S. *et al.* Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16,** 906–914 (2000).
3. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* **98,** 13790–13795 (2001).
4. Shipp, M. A. *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* **8,** 68–74 (2002).
5. Huang, D.-S. & Zheng, C.-H. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* **22,** 1855–1862 (2006).
6. Ghosh, D. & Chinnaiyan, A. M. Classification and selection of biomarkers in genomic data using LASSO. *J Biomed Biotechnol* **30,** 147–154 (2005).
7. Chen, S., Donoho, D. & Saunders, M. Atomic Decomposition by Basis Pursuit. *SIAM Review* **43,** 129–159 (2001).
8. Candes, E. J., Romberg, J. & Tao, T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on* **52,** 489–509 (2006).
9. Candes, E. J. & Tao, T. Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? *Information Theory, IEEE Transactions on* **52,** 5406–5425 (2006).
10. Donoho, D. L. Compressed sensing. *Information Theory, IEEE Transactions on* **52,** 1289–1306, (2006).
11. Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S. & Yi, M. Robust Face Recognition via Sparse Representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **31,** 210–227 (2009).
12. Kim, S. J., Koh, K., Lustig, M., Boyd, S. & Gorinevsky, D. An Interior-Point Method for Large-Scale l_1 -Regularized Least Squares. *IEEE Journal of Selected Topics in Signal Processing* **1,** 606–617 (2007).
13. Hang, X. & Wu, F.-X. Sparse Representation for Classification of Tumors Using Gene Expression Data. *Journal of Biomedicine and Biotechnology* **2009,** 6 (2009).
14. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* **101,** 4164–4169, doi:10.1073/pnas.0308531101 (2004).
15. Chun-Hou, Z. Metasample-Based Sparse Representation for Tumor Classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8,** 1273–1282 (2011).
16. Liebermeister, W. Linear modes of gene expression determined by independent component analysis. *Bioinformatics* **18,** 51–60 (2002).
17. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* **97,** 10101–10106 (2000).
18. Liao, B. *et al.* Learning a weighted meta-sample based parameter free sparse representation classification for microarray data. *PLoS One* **9,** e104314 (2014).
19. Candès, E. J., Wakin, M. B. & Boyd, S. P. Enhancing Sparsity by Reweighted ℓ 1 Minimization. *Journal of Fourier Analysis and Applications* **14,** 877–905, doi:10.1007/s00041-008-9045-x (2008).
20. Belhumeur, P. N., Hespanha, J. P. & Kriegman, D. J. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19,** 711–720 (1997).
21. Bartlett, M. S., Movellan, J. R. & Sejnowski, T. J. Face recognition by independent component analysis. *IEEE Transactions on Neural Networks* **13,** 1450–1464 (2002).
22. Wang, S. L., Zhu, Y. H., Jia, W. & Huang, D. S. Robust Classification Method of Tumor Subtype by Using Correlation Filters. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9,** 580–591, (2012).
23. Seung-Jean, K., Koh, K., Lustig, M., Boyd, S. & Gorinevsky, D. An Interior-Point Method for Large-Scale l1-Regularized Least Squares. *Selected Topics in Signal Processing, IEEE Journal of* **1,** 606–617, (2007).
24. Zheng, C. H., Huang, D. S., Zhang, L. & Kong, X. Z. Tumor Clustering Using Nonnegative Matrix Factorization With Gene Selection. *IEEE Transactions on Information Technology in Biomedicine* **13,** 599–607(2009).
25. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* **96,** 6745–6750 (1999).
26. Nutt, C. L. *et al.* Gene Expression-based Classification of Malignant Gliomas Correlates Better with Survival than Histological Classification. *Cancer Research* **63,** 1602–1607 (2003).
27. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403,** 503–511 (2000).
28. Khan, J. *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* **7,** 673–679 (2001).
29. Yeoh, E.-J. *et al.* Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1,** 133–143 (2002).
30. Armstrong, S. A. *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* **30,** 41–47 (2002).
31. Robnik-Šikonja, M. & Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning* **53,** 23–69 (2003).
32. Amaldi, E. & Kann, V. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theor. Comput. Sci.* **209,** 237–260 (1998).

33. Boyd, S. & Vandenberghe, L. *Convex Optimization* (Cambridge University Press, 2004).
34. Li, Z. *et al.* Kernel Sparse Representation-Based Classifier. *Signal Processing, IEEE Transactions on* **60,** 1684–1695,(2012).

Acknowledgements
This study is supported by the Program for New Century Excellent Talents in university (Grant No.NCET-10-0365), National Nature Science Foundation of China (Grant Nos 11171369, 61272395, 61370171, 61300128, 61472127, 61572178 and 61672214).

Author Contributions
W.B.L. conceived the project, developed the main method, designed and implemented the experiments, analyzed the result, and wrote the paper. B.L., W.Z. analyzed the result, and wrote the paper. M.C., L.P., X.H.W., C.L.G. implemented the experiments, and analyzed the result. K.Q.L. analyzed the result. All authors reviewed the final manuscript.

Additional Information
Supplementary information accompanies this paper at http://www.nature.com/srep

Competing Interests: The authors declare no competing financial interests.

How to cite this article: Li, W. *et al.* Maxdenominator Reweighted Sparse Representation for Tumor Classification. *Sci. Rep.* **7**, 46030; doi: 10.1038/srep46030 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2017