SCIENTIFIC REPORTS

OPEN

Received: 23 March 2018
Accepted: 31 May 2018
Published online: 14 June 2018

Fisher Discrimination Regularized Robust Coding Based on a Local Center for Tumor Classification

Weibiao Li[1], Bo Liao[1], Wen Zhu[1], Min Chen[1], Zejun Li[1], Xiaohui Wei[1], Lihong Peng[2], Guohua Huang[1], Lijun Cai[1] & HaoWen Chen[1]

Tumor classification is crucial to the clinical diagnosis and proper treatment of cancers. In recent years, sparse representation-based classifier (SRC) has been proposed for tumor classification. The employed dictionary plays an important role in sparse representation-based or sparse coding-based classification. However, sparse representation-based tumor classification models have not used the employed dictionary, thereby limiting their performance. Furthermore, this sparse representation model assumes that the coding residual follows a Gaussian or Laplacian distribution, which may not effectively describe the coding residual in practical tumor classification. In the present study, we formulated a novel effective cancer classification technique, namely, Fisher discrimination regularized robust coding (FDRRC), by combining the Fisher discrimination dictionary learning method with the regularized robust coding (RRC) model, which searches for a maximum a posteriori solution to coding problems by assuming that the coding residual and representation coefficient are independent and identically distributed. The proposed FDRRC model is extensively evaluated on various tumor datasets and shows superior performance compared with various state-of-the-art tumor classification methods in a variety of classification tasks.

Microarray techniques have been used to delineate cancer groups or to identify candidate genes for cancer prognosis. The accurate classification of tumors is important for cancer treatment. With the advancement of DNA microarray and next-generation sequencing technology[1–4], various gene expression profile (GEP) data are rapidly obtained. Thus, we should develop novel analysis methods that can deeply mine and interpret these data to obtain insight into the mechanisms of tumor development. To date, a number of methods have been proposed for classifying cancer types or subtypes[5–9]. These common methods, including support vector machine[10], linear discriminant analysis[11], partial least squares (PLS)[12], and artificial neural networks[13], have been used to mine gene expression data.

Machine learning-based methods have been widely used in tumor classification. However, these methods require a predictive model to predict the labels of test samples. Predictive model selection is a complex training procedure that easily leads to overfitting and decreased prediction performance. Recently, given the non-requirement for model selection and robustness to noise, outliers, and incomplete measurements, sparse representation-based classifier (SRC) was proposed for face recognition[14,15] and further extended to cancer classification[16–18] and miRNA-disease association prediction[19,20]. For example, Hang *et al.* proposed a SRC-based method to classify six tumor gene expression datasets and obtained excellent performance[18]. Zheng *et al.* further combined the idea of metasample and proposed a new SRC-based method for tumor classification called metasample-based sparse representation-based classifier (MSRC)[16]. These experiments showed that MSRC is efficient for tumor classification and can achieve high accuracy. Li *et al.* proposed a new classifier called the max-denominator reweighted sparse representation-based classifier (MRSRC) for cancer classification[5]. These experiments showed the efficiency and robustness of MRSRC. All SRC-based methods model a classification problem to identify a sparse representation of test samples, whereas the L1 sparsity constraint represents a test sample as the linear combination of these training samples.

In the sparse representation model, the test sample $y \in R^m$ is used to represent a dictionary $D = \{D_1, D_2, \ldots D_c\} \in R^{m\times n}$, that is, $y \approx D\alpha$ where the sparse representation vector $\alpha \in R^n$ only shows several large entries. Then, the

[1]College of Information Science and Engineering, Hunan University, Changsha, Hunan, 410082, China. [2]Hunan University of Technology, Zhu Zhou, Hunan, 412007, China. Correspondence and requests for materials should be addressed to B.L. (email: dragonbw@163.com)

test samples are classified based on the solved vector αand the dictionary D. The selection of vector α and the dictionary D is crucial to the success of the sparse representation model. The previously described SRC-based methods directly regarded the training samples of all classes as the dictionary to represent the test sample and classified the test sample by evaluating which class leads to minimal reconstruction error. Although these methods showed interesting results, noise, outliers, incomplete measurements, and trivial information in the raw training data made this classification less effective. These naive methods also do not make maximize the discriminative information in the training samples. These problems can be addressed by properly learning a discriminative dictionary.

In general, discriminative dictionary learning methods can be divided into two categories. In the first category, a dictionary shared by all classes is learned, whereas the representation coefficients are discriminative. Jiang *et al*. proposed that samples of the same class possesses similar sparse representation coefficients[21]. Mairal *et al*. proposed a task-driven dictionary learning framework that minimizes the different risk functions of the representation coefficients for different tasks[22]. In general, these of methods aims to learn a shared dictionary by all classes and classify test samples with representation coefficients. However, the shared dictionary loses the class labels of the dictionary atoms. Thus, classifying the test samples based on the class-specific representation residuals is not feasible.

In the second category, discriminative dictionary learning methods learn a dictionary class by class, and atoms of the dictionary correspond to the subject class labels. Yang *et al*. learned a dictionary for each class, classified the test samples by using the representation residual, and applied dictionary learning methods to face recognition and signal clustering[23]. Wang *et al*. proposed a class-specific dictionary learning method for sparse modeling in action recognition[24]. In the previously mentioned methods, test samples are classified by using the representation residual associated with each class, but the representation coefficients are not used and are not enforced to be discriminative in the final classification.

To solve the previously discussed problems, Yang *et al*. proposed a Fisher discrimination dictionary learning framework to learn a structured dictionary[25]. In discrimination dictionary learning, the sparse representation coefficients present large between-class scatter and small within-class scatter. Each class-specific sub-dictionary presents good reconstruction of the training samples from that class and poor reconstruction of the other classes. By Fisher discrimination dictionary learning, the representation residual associated with each class can effectively be used for classification and the discrimination of representation coefficients can be exploited.

All SRC-based methods assume that the coding residual follows a Gaussian or Laplacian distribution, which may not be effective for describing the coding residual in practical GEP datasets. To address this problem, Yang *et al*. proposed a regularized robust coding (RRC) method for face recognition[26]. The RRC model searches for a maximum a posteriori (MAP) solution of the coding problem by assuming that the coding residual and representation coefficient are independent and identically distributed. However, either SRC-based or RRC methods or both do not take full advantage of discriminative information in representation coefficients. In the present study, we present RRC based on the Fisher discrimination dictionary learning method, a novel and effective cancer classification technique combining RRC methods and the concept of Fisher discrimination dictionary learning, which can maximize the use of discriminative information in representation coefficients and representation residuals. The proposed Fisher discrimination regularized robust coding (FDRRC) model extensively applies to various tumor GEP datasets and shows superior performance to different state-of-the-art SRC-based and machine learning-based methods in a variety of classification tasks.

The remainder of the paper is organized as follows: Section 2 mainly describes the experimental process and presents the experimental results obtained from eight tumor datasets. Section 3 discusses the proposed method, concludes the paper and outlines future studies. Section 4 describes the fundamentals of FDRRC.

Results

In present study, eight publicly available tumor data sets are used to evaluate the performance of FDRRC. The experiment is divided into four sections. In the first section, cancer datasets and dataset preprocessing are introduced. In the second section, parameter selection is discussed. In the third section, describes the various samples used in the experiment with 400 top genes on eight datasets. In the fourth section, to make a fair performance comparison, cross-validation (CV) is presented. The proposed method is compared with several representative methods, such as SRC[18], SVD + MSRC[27] and MRSRC[5]. SRC, MSRC, and MRSRC are SRC-based methods that have been widely used in tumor classification in recent years. All experiments are implemented in the Matlab environment and conducted on a personal computer (Intel Core dual-core CPU with 2.93 GHz and 8 G RAM).

Cancer datasets and dataset preprocessing.

For a more comprehensive comparison of the performance of these methods, eight tumor GEP datasets are used to evaluate the proposed method. These datasets include five two-class datasets and three multi-class datasets. The summarized descriptions of the eight GEP datasets are provided in Table 1.

The five two-class tumor datasets are acute leukemia dataset[28], colon cancer dataset[29], gliomas dataset[30], diffuse large B-cell lymphoma (DLBCL) dataset[31] and Prostate dataset[32]. The acute leukemia set contains 72 samples from two subclass. The colon cancer data set includes 62 samples, with gene expression data for 40 tumor and 22 normal colon tissue samples. The gliomas data set consists of 50 samples from two subclasses (glioblastomas and anaplastic oligodendrogliomas), and each sample contains 12,625 genes. For the DLBCL data set, RNA was hybridized to high-density oligonucleotide microarrays to measure the gene expression. The target dataset 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. For the prostate tumor data set, the gene expression profiles were derived from tumors and non-tumor samples from prostate cancer patients, including 59 normal and 75 tumor samples. The number of genes is 12,600. Table 1 provides the details of the data sets.

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
Prostate data	2	12,600	136
ALL data	6	12,625	248
MLLLeukemia data	3	12,582	72
LukemiaGloub data	3	7,129	72

Table 1. The descriptions of eight data sets of tumor.

Dataset	SRC	MSRC	MRSRC	FDRRC
Colon cancer data	77.42	80.65	82.26	**83.87**
Acute leukemia data	94.44	95.83	95.83	**98.61**
Gliomas data	70.00	70.00	74.00	**82.00**
DLBCL data	90.91	92.21	89.61	**96.10**
Prostate data	88.24	95.10	**96.08**	92.16
ALL data	97.18	97.58	**97.98**	**97.98**
MLLLeukemia data	97.22	**98.61**	**98.61**	**98.61**
LukemiaGloub data	94.44	95.83	97.22	**100**

Table 2. 10-fold CV prediction accuracy of eight tumor microarray datasets by using various classification methods with the top 400 genes.

For multi-class datasets, the data sets include the small round blue cell tumors (ALL)[33], MLLLeukemia[34], and LukemiaGloub[28]. The ALL data set total contains 248 samples and 12,626 genes from six subclasses. The MLLLeukemia data set contains 72 samples and 12,582 genes per sample with three subclasses. The LukemiaGloub data set contains 72 samples with three subclasses. Each sample contains 7,129 genes. Table 4 provides details of the data sets.

GEP data offer high dimensionality and a small sample size. Redundant and irrelevant data significantly affects classification. To compare the performance of FDRRC and SRC-based methods in the gene selection, the ReliefF algorithm is applied to the training set[35]. Then, the top 400 genes are selected from each dataset, thereby presenting a good trade-off between computational complexity and biological significance.

Parameter selection. Five parameters should be set in the FDRRC model. The dictionary learning phase employs two parameters: λ_1 and λ_2, which are both presented in Eq.(8). In general, we search λ_1, λ_2 from a small set {0.001, 0.005, 0.01, 0.05, 0.1} by five-fold CV. The classifying phase includes three parameters, namely, μ and δ from the weight function Eq. (21) and w from residual function Eq. (24). Parameter μ controls the decreased rate of the weight $w_{i,j}$; we can simply set $\mu = s/\delta$, where $s = 8$ is a constant. Parameter δ controls the location of the demarcation point, which can be obtained by using the following formula:

$$\delta = \pi(e)_\varphi, \tag{1}$$

where $\pi(e)_\varphi$ is the φ^{th} largest element of the set $\left\{e_j^2,\ j = 1,\ 2,\ \cdots,\ m\right\}$ and $\varphi = \varsigma(\tau m)$ outputs the largest integer smaller than τm. According to the experiments[7], $\tau = 0.9$ can be set in the classification of tumors. Parameter w can balance the contributions of the representation residual and representation vector to the classification. We search for w from a small set {0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1} by five-fold VC.

Comparison of the balance division performance. Different divisions of the training set and test set can greatly affect the classification performance. To avoid the effects of an imbalanced training set, the balance division method (BDM) is designed to divide each original data set into a balanced training set and test set. For this BDM, Q samples from each subclass are randomly selected for use in the training set, and the remaining samples are used in the test set. Here, Q is an integer number. In the present study, we set Q = 5to $\min(|c_i|) - 1$ samples per subclass as the training set and used the remaining samples for testing to guarantee that at least one sample in each category can be used in the test. Q denotes the number of training samples per class, and $\min(|c_i|)$ denotes the minimum number of subclass set of samples in the training data. Suggesting that when Qis 5, then 5 samples per-subclass are randomly selected and used as the training set and the rest are assigned to the test set. In this experiment, the training/testing is performed 10 times, and the average classification accuracies are presented.

The average prediction accuracies that vary with different values of Q are shown in Figs 1 and 2, showing that, in the case of two-class classification, FDRRC achieves the highest classification accuracy in most cases in the acute leukemia and Gliomas datasets. Although gliomas are difficult to classify, FDRRC can still achieve the

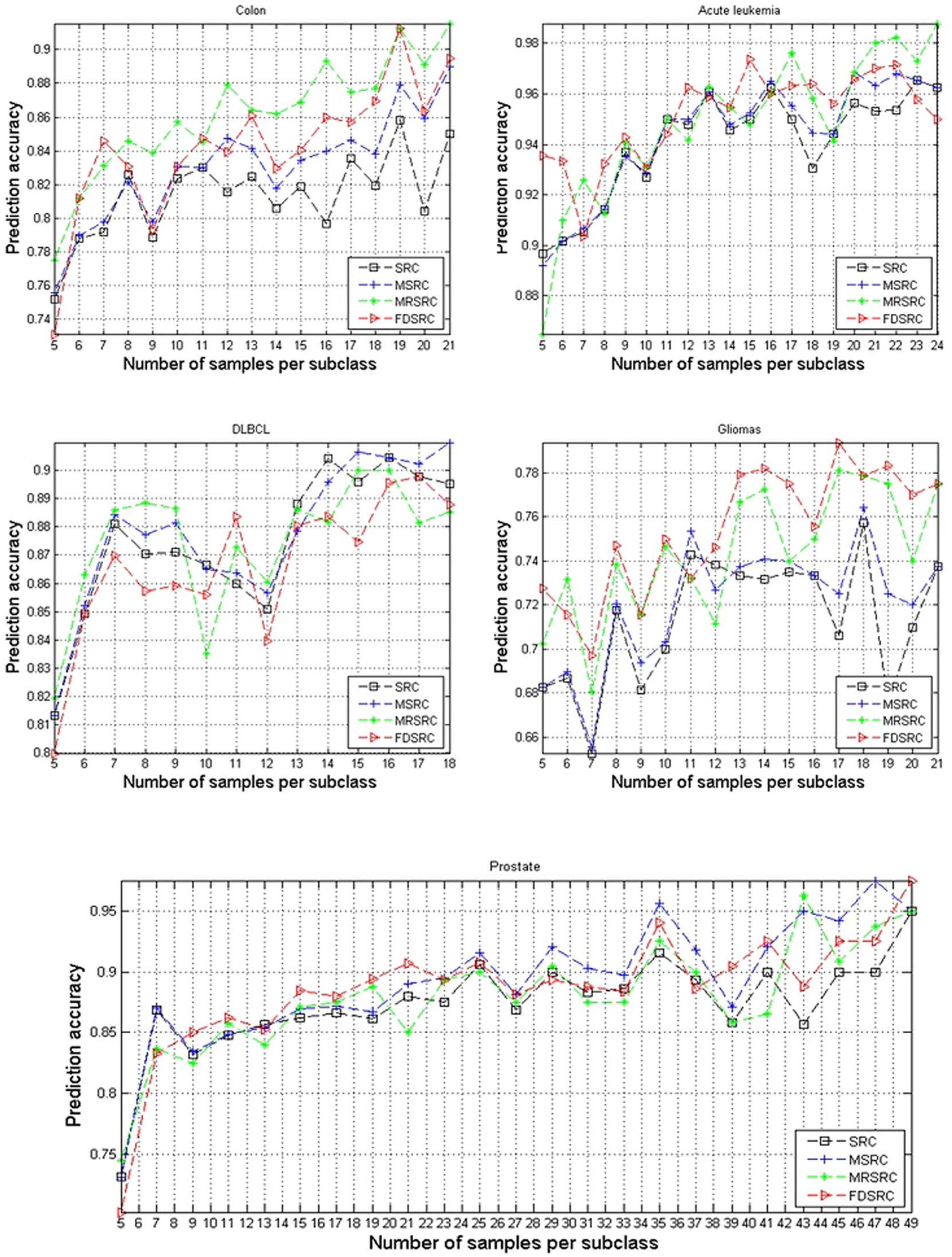

Figure 1. Comparison of prediction accuracy on five two-class classification datasets by varying the number of samples from per subclass.

highest classification accuracy when Q = 17 samples per subclass are used in training. For the prostate dataset, FDRRC achieves the highest classification accuracy in most cases when the samples are few per subclass. In the case of multi-class classification, the experimental results indicate that FDRRC obtains a significant advantage in the ALL and MLLLeukemia datasets. Generally, the present methods are superior to other SRC-based methods in prediction accuracy not only on the four two-class classification datasets but also on the three multi-class classification datasets.

Comparison with different numbers of genes. To compare the performance of the four models with different feature dimensions on eight tumor data sets, we run experiments using the ReliefF algorithm to select genes from 10^2 to 30^2 in increments of 5. For these experiments, the number of samples per subclass of the training set, was selected from {5, 6, 7, 8, 9, 10} by five-fold VC. The results are shown in Fig. 3.

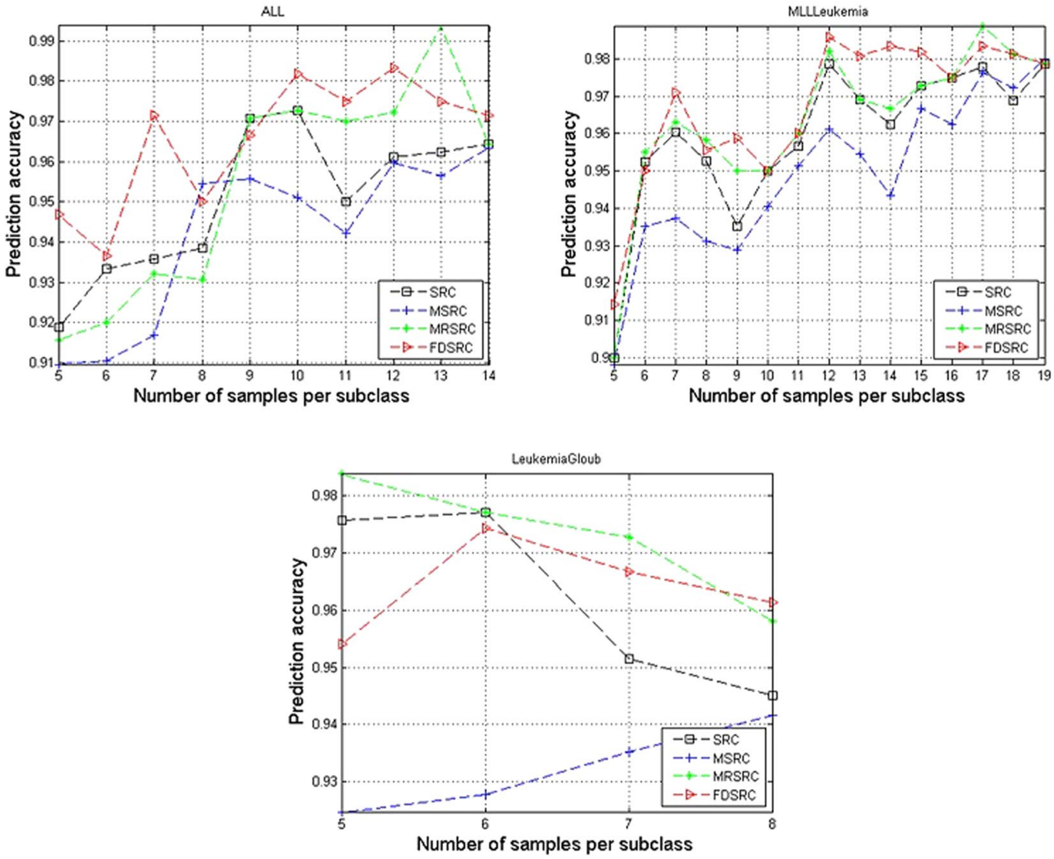

Figure 2. Comparison of prediction accuracy on three multi-class classification datasets by varying the number of samples from per subclass.

Figure 3 presents the average prediction accuracy for the classification of eight tumor data sets. As shown in Fig. 3, FDRRC achieves the best accuracy in the five data sets in most cases, illustrating that FDRRC is robust with respect to the number of top genes. For Colon, Acute leukemia, DLBCL, Gliomas, Prostate and MLLLeukemia data sets, the accuracy of the curve increases with the increasing number of genes selected. Clearly, the selection of the top genes can improve the performance of all classification methods. For Acute leukemia dataset and ALL dataset, the best number of top genes is 400. These results suggest that the selection of the top 400 genes is reasonable.

Comparison of 10-fold CV performance.

To evaluate the classification performance on imbalanced split training/testing sets, we perform a 10-fold stratified CV experiment to evaluate the classification performance between FDRRC and SRC-based methods. All samples are randomly divided into 10 subsets and nine subsets are used for training, the remaining samples are used for testing.

The 10-fold CV results are summarized in Tables 2, 3 and 4. Table 2 shows that FDRRC achieves the highest level of accuracy in seven datasets. Particularly in multi-class datasets, FDRRC exhibits the best classification accuracy in all datasets. Table 3 indicates that FDRRC achieves the highest prediction sensitivity in six datasets, whereas FDRRC shows the best classification sensitivity in four tow-class datasets. Table 4 shows that FDRRC exhibits the highest specificity in seven datasets. Particularly in multi-class datasets, FDRRC exhibits the best classification accuracy in all datasets. Thus, we concluded that the excellent applicability of FDRRC whether in two-class or multi-class datasets, exhibits the best classification accuracy, the best classification sensitivity, and the best classification specificity in most cases.

Discussions

The results of the present study, show that FDRRC outperforms the sparse representation-based methods (such as SRC, MSRC, and MRSRC) in most experiments. FDRRC outperforms the sparse representation-based methods probably because the representation residual associated with each class can be effectively used for classification, the discrimination of representation coefficients has been exploited, the coding residual is independent and identically distributed and the local center can help to distinguish outliers.

In the present, we proposed a new method, called FDRRC for classifying tumors. This method adopts the Fisher discrimination dictionary learning method and the concept of the local center with the RRC model. The FDRRC model learns a discriminative dictionary and seeks a MAP solution to the coding problem. Classification is achieved by a local center classifier, which takes full discriminative information in representation coefficients.

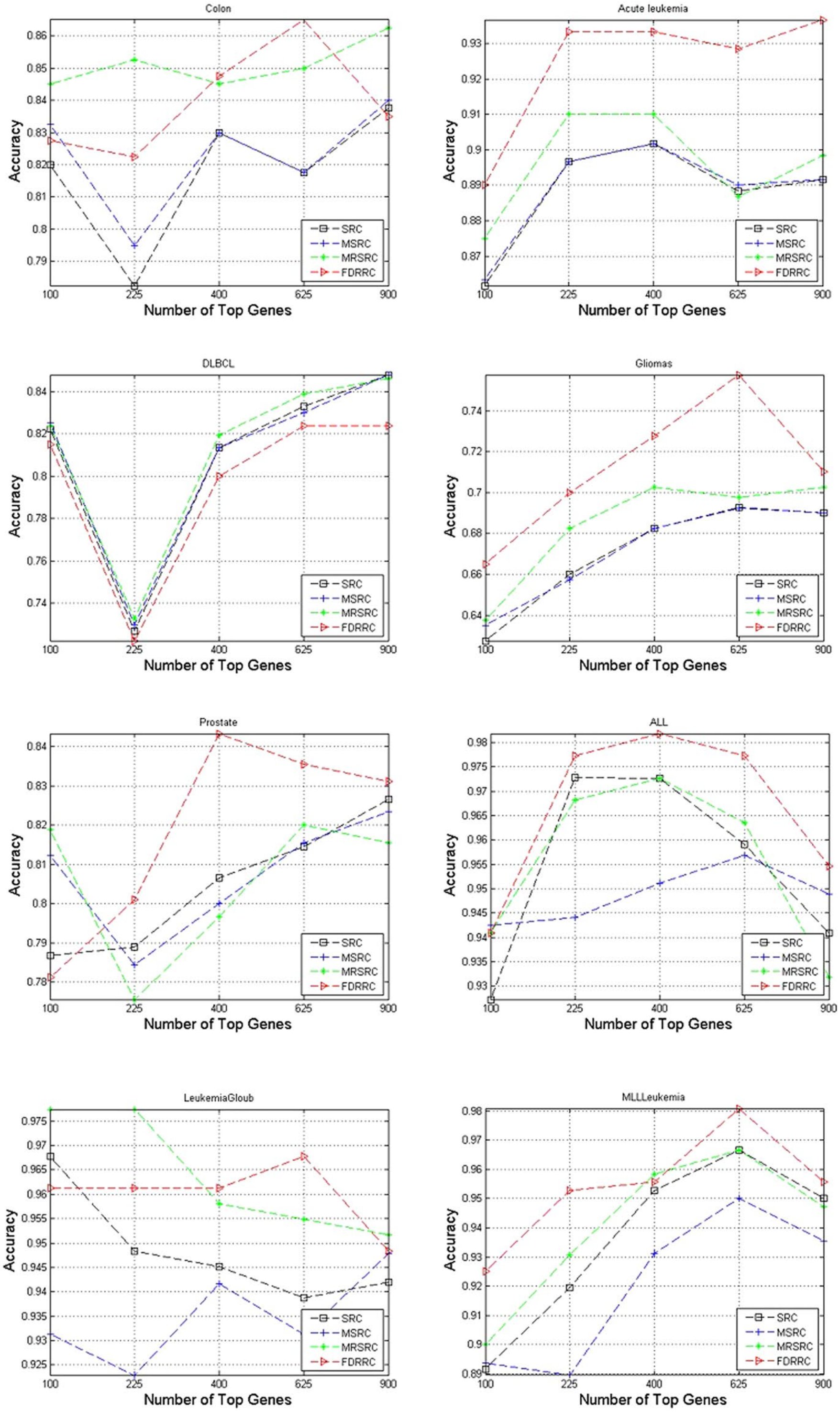

Figure 3. Comparison of accuracy on eight datasets by varying the number of top selected genes.

We also compare the performance of FDRRC with those of three sparse representation-based methods by using eight tumor expression datasets. The results demonstrate the superiority of FDRRC and validate the effectiveness and efficiency of FDRRC in tumor classification.

Compared with the other methods, FDRRC exhibits a stable performance with respect to various datasets. The properties of this FDRRC algorithm should be further investigated. Thus, we will extend the algorithm with a superior discriminative dictionary and consider the driver genes to tailor the algorithm in our future studies. In addition, FDRRC will be used to predict miRNA[36] and lncRNA-disease association[37] in future studies.

Dataset	SRC	MSRC	MRSRC	FDRRC
Colon cancer data	68.18	68.18	**81.82**	77.27
Acute leukemia data	92.00	92.00	88.00	**96.00**
Gliomas data	71.43	71.43	71.43	**82.14**
DLBCL data	94.74	94.74	**100**	**100**
Prostate data	92.31	94.23	**96.15**	**96.15**
ALL data	80.00	86.67	**93.33**	86.67
MLLLeukemia data	95.83	**100**	**100**	**100**
LukemiaGloub data	88.89	88.89	88.89	**100**

Table 3. 10-fold CV prediction sensitivity of eight tumor microarray datasets by using various classification methods with the top 400 genes.

Dataset	SRC	MSRC	MRSRC	FDRRC
Colon cancer data	82.50	**87.50**	82.50	**87.50**
Acute leukemia data	95.74	97.87	**100**	**100**
Gliomas data	68.18	68.18	77.27	**81.82**
DLBCL data	89.66	91.38	86.21	**94.83**
Prostate data	84.00	**96.00**	**96.00**	88.00
ALL data	99.14	98.71	98.71	**99.57**
MLLLeukemia data	**100**	**100**	**100**	**100**
LukemiaGloub data	**100**	**100**	**100**	**100**

Table 4. 10-fold CV prediction specificity of eight tumor microarray datasets by using various classification methods with the top 400 genes.

Methods

Description of SRC problem. Assuming that X = $\{X_1, X_2, \ldots, X_c\} \in R^{m\times n}$ is a training sample set, where c corresponds to the number of subclasses, and m, n are dimensionality and the number of samples, respectively. The j_{th} class training samples X_j can be presented as columns of a matrix $X_j = [x_{j,1},\ x_{j,2},\ \cdots x_{j,n_j}] \in R^{m\times n},\ j = 1,\ 2,\ \cdots,\ c$ where $x_{j,i}$ is a sample of j_{th} class, and n_j refers to the number of j_{th} class training samples. Let L = $\{l_1, l_2, \ldots l_c\}$ denote the label set, whereas y $\in R^m$ is a test sample. Then, the SRC-based problem can be represented as follows:

$$\alpha^{\wedge} = argmin_{\alpha}\{\|y - X\alpha\|_2^2 + \gamma\|\alpha\|_1\} \tag{2}$$

where $\alpha^{\wedge} = [\alpha_1^{\wedge},\ \alpha_2^{\wedge},\ \cdots,\ \alpha_c^{\wedge}]$ includes the sparse representation coefficient of y with respect to X, and γ is a small positive constant. By obtaining representation coefficient $\alpha^{\wedge}$, SRC-based method assigns a label to test sample y according to the following equation:

$$e_i = \|y - X_i\alpha_i^{\wedge}\|_2^2 \tag{3}$$

where $\alpha_i^{\wedge}$ is the sparse representation coefficient sub-vector associated with subclass X_i. The classification rule is set as $identity(y) = argmin_i\{e_i\}$.

Fisher Discrimination Dictionary Learning. Given the training samples X = $\{X_1, X_2, \ldots, X_c\}$, the Fisher discrimination dictionary learning model not only requires that D should be highly capable of representing X (i.e., X ≈ Dα) but also that D can strongly distinguish the samples in X. The Fisher discrimination dictionary learning model can be expressed as follows:

$$J_{(D,X)} = argmin_{(D,X)}\left\{r(X,\ D,\ \alpha) + \lambda_1\|\alpha\|_1 + \lambda_2 f(\alpha)\ s.\ t.\ \ \|d_n\|_2 = 1,\ \forall n\right\} \tag{4}$$

where $f(\alpha)$ is a discrimination term imposed on the coefficient matrix α, $\|a\|_1$ is the sparsity penalty, $r(X, D, \alpha)$is the discriminative data fidelity term, and λ_1 and λ_2 are scalar parameters.

We can write α_i as $\alpha_i = [\alpha_i^1;\ \cdots;\ \alpha_i^j;\ \cdots;\ \alpha_i^c]$, where α_i^j is the representation coefficient of X_i over D_i. For the discriminative data fidelity term $r(X, D, \alpha)$, X_i could be well represented by D_i but not by $D_j j \neq i$. This relationship indicates that α_i^i should present several significant coefficients to achieve a small $\|X_i - D_i\alpha_i^i\|_F^2$, whereas α_i^j, $j \neq i$ should include small coefficients so that $\|D_i\alpha_i^i\|_F^2$ is small. Thus, the discriminative data fidelity term can be defined as follows:

Input: $\sigma,\ \tau > 0$.
1. Initialization: $\alpha_i^{\wedge(1)} = 0$ and h = 1.
2. while convergence or the maximal itertion number is not reached do h + h = 1
$\alpha_i^{\wedge(h)} = S_{\tau/\sigma}\left(\alpha_i^{\wedge(h-1)} - \frac{1}{2\sigma}\nabla Q(\alpha_i^{\wedge(h-1)})\right)$
where $\nabla Q(\alpha_i^{\wedge(h-1)})$ is the derivative of $Q(\alpha_i)$ w.r.t $\alpha_i^{\wedge(h-1)}$, and $S_{\tau/\sigma}$ is a component-wise soft thresholding operator defined by Wright *et al.*[42].
$[S_{\tau/\sigma}(\alpha)]_j = \begin{cases} 0 & \lvert\alpha_j\rvert \le \tau/\sigma \\ \alpha_j - sign(\alpha_j)\tau/\sigma & otherwise \end{cases}$
3. Return $\alpha_i^{\wedge} = \alpha_i^{\wedge(h)}$.

Table 5. Update of representation coefficient α in the Fisher discrimination dictionary learning model.

$$r(X_i, D, \alpha_i) = \|X_i - D\alpha_i\|_F^2 + \|X_i - D_i\alpha_i^{\,i}\|_F^2 + \sum_{\substack{j=1\\ j\neq i}}^{c}\left\|D_j\alpha_i^{\,j}\right\|_F^2. \tag{5}$$

For the discriminative coefficient term *f*(α), the Fisher discrimination criterion[38] is expected to minimize the within-class scatter of α, denoted by *SW*(α), and maximize the between-class scatter of α, denoted by *SB*(α). *SW*(α) and *SB*(α) are defined as follows:

$$\mathrm{SW}(\alpha) = \sum_{i=1}^{c}\sum\nolimits_{\alpha_c\in\alpha_i}(\alpha_c - m_i)(\alpha_k - m_i)^T \ and\ SB(\alpha) = \sum_{i=1}^{c} n_i(m_c - m)(m_i - m)^T, \tag{6}$$

where m_i and *m* are the mean vectors of α_i and α, respectively, and n_i is the number of samples in class X_i. Thus, the criminative coefficient term can be defined as follows:

$$f(\alpha) = tr(\mathrm{SW}(\alpha)) - tr(\mathrm{SB}(\alpha)) + \eta\|\alpha\|_F^2 \tag{7}$$

where $tr(\cdot)$ means the trace of a matrix, η is a parameter, and $\|\alpha\|_F^2$ is an elastic term.

Finally, the Fisher discrimination dictionary learning model can be expressed as follows:

$$min_{(D,X)}\left\{\sum_{i=1}^{c} r(X_i, D, \alpha_i) + \lambda_1\|\alpha\|_1 + \lambda_2(tr(\mathrm{SW}(\alpha)) - tr(\mathrm{SB}(\alpha))) + \eta\|\alpha\|_F^2\right\} s.\ t.\ \ \|d_n\|_2 = 1,\ \forall n \tag{8}$$

Optimization of the Fisher discrimination dictionary learning model can be divided into sub-problems, that is, updating α with a fixed D and updating D with a fixed α.

When α is updated, the dictionary D is fixed and can compute α_i class by class. When computing α_i, all α_j, $j \neq i$ are fixed. The objective function expressed in Eq. (8) is reduced to a sparse representation problem and can be written as follows:

$$min_{\alpha_i}\left\{r(X_i, D, \alpha_i) + \lambda_1\|\alpha_i\|_1 + \lambda_2 f_i(\alpha_i)\right\} \tag{9}$$

with

$$f_i(\alpha_i) = \|\alpha_i - M\|_F^2 - \sum_{k=1}^{c}\|M_k - M\|_F^2 + \eta\|\alpha_i\|_F^2,$$

where M_k and *M* are the mean vector matrices of class k and all classes, respectively. In this study, we set $\eta = 1$ for simplicity. Notably, all terms in Eq. (9), except for $\|a\|_1$, are differentiable. We rewrite Eq. (9) as follows:

$$min_{\alpha_i}\left\{Q(\alpha_i) + 2\tau\|\alpha_i\|_1\right\}, \tag{10}$$

where $Q(\alpha_i) = r(X_i, D, \alpha_i) + \lambda_2 f_i(\alpha_i)$ and $\tau = \lambda_1/2$. The method of FISTA[39] can be employed to solve Eq. (10), as described in Table 5.

When updating $D = [D_1, D_2, \ldots, D_c]$, the coefficient α is fixed. We also update $D_i = \left[d_1,\ d_2,\ \cdots,\ d_{n_i}\right]$ class by class. When updating D_i, all D_j, $j \neq i$, are fixed. The objective function expressed in Eq. (8) is reduced to:

$$min_{D_i}\left\{\|X^{\sim} - D_i\alpha^i\|_F^2 + \|X_i - D_i\alpha_i^{\,i}\|_F^2 + \sum_{j=1, j\neq i}^{c}\left\|D_i\alpha_j^{\,i}\right\|_F^2\right\} s.\ t.\ \ \|d_l\| = 1,\ l = 1, 2, \cdots, n_i \tag{11}$$

where $X^{\sim} = X - \sum_{j=1,j\neq i}^{c} D_j\alpha^j$ and α^j is the representation matrix of *X* over D_i. Eq. (11) could be re-written as follows:

Fix α and update each $D_i,\ \ i = 1,\ 2,\ \ldots C$, by solving Eq. (12)
1. Let $Z_i = [z_1;\ z_2;\ \ldots;\ z_{n_i}]$ and $D_i = \left[d_1,\ d_2,\ \ldots d_{n_i}\right]$, where z_j, $j = 1, 2, \ldots n_i$ is the row vector of z_i, and d_j is the j_{th} column vector of D_i.
2. Fix all $d_j,\ \ l \neq j$, update d_j. Let $Y = \Lambda_i - \sum_{l \neq j} d_l z_l$. The minimization of Eq. (12) becomes
$\min_{d_j} \left\| Y - d_j z_j \right\|_F^2 s.\ t.\ \left\| d_j \right\|_2 = 1$
After some deviation, we could get the solution $d_j = Y z_j^T / \left\| Y z_j^T \right\|_2$.
3. Then Fix D and update α like Table 5.

Table 6. Update of dictionary D in the Fisher discrimination dictionary learning model.

$$min_{D_i} \|\Lambda_i - D_i Z_i\|_F^2\ s.\ t.\ \ \|d_l\|_2 = 1,\, l = 1,\, 2,\ \cdots,\ n_i \tag{12}$$

where $\Lambda_i = [X^{\sim}\ X_i 0\ \ldots\ 00\ \ldots\ 0]$, $Z_i = \alpha^i \alpha_i^i \alpha_1^i \cdots \alpha_{i-1}^i \alpha_{i+1}^i \cdots \alpha_c^i$ and 0 is a zero matrix with the appropriate size based on the context. Eq. (12) can be efficiently solved by updating each dictionary atom one by one via the algorithm of Yang *et al.*[40]. The update of dictionary D is described in Table 6.

Description of RRC.

In the SRC-based method, coding residual $e = y - D\alpha$ follows Gaussian distribution[25]. However, in practice, Gaussian priors on e may be invalid, especially when GEP data are corrupted and contain outliers. To deal with this problem, we can consider tumor classification from the view point of Bayesian estimation, especially MAP estimation. Based on MAP estimation, sparse representation coefficient α can be expressed as follows[26]:

$$\alpha^{\wedge} = argmax_{\alpha}\ \ln p(\alpha|y) \tag{13}$$

Then, by using Bayesian formulation, we can obtain the following:

$$\alpha^{\wedge} = argmax_{\alpha}\{\ln p(y|\alpha) + \ \ln p(\alpha)\} \tag{14}$$

Assuming that elements e_i of coding residual $e = y - D\alpha = [e_1; e_2; \ldots e_m]$ are independent and identically distributed and feature the probability density function (PDF) $f_\theta(e_i)$, then we can obtain the equation below:

$$\ln p(y|\alpha) = \prod_{i=1}^{m} f_\theta(y_i - r_i\alpha) \tag{15}$$

Meanwhile, assuming that element α_i of sparse representation coefficient $\alpha = [\alpha_1; \alpha_2; \ldots; \alpha_n]$ are independent and identically distributed and contain the PDF $f_\sigma(\alpha_i)$, then we can acquire the following formula:

$$\mathrm{p}(\alpha) = \prod_{j=1}^{n} f_\sigma(\alpha_j) \tag{16}$$

Finally, MAP estimation of α can be expressed as follows:

$$\alpha^{\wedge} = argmax_{\alpha}\left\{\prod_{i=1}^{m} f_\theta(y_i - r_i\alpha) + \prod_{j=1}^{n} f_\sigma(\alpha_j)\right\} \tag{17}$$

Letting $\rho_\theta(e) = -\ln f_\theta(e)$ and ${}_{\rho\sigma}(\alpha) = -ln f_\sigma(\alpha)$, then, the above equation can be converted into the following:

$$\alpha^{\wedge} = argmax_{\alpha}\left\{\sum_{i=1}^{m} \rho_\theta(y_i - r_i\alpha) + \sum_{j=1}^{n} \rho_\sigma(\alpha_j)\right\} \tag{18}$$

The above model is called RRC. Two key issues must be considered to solve the RRC model: determining distributions of $\rho_\theta(e)$ and $\rho_\sigma(\alpha)$; and minimizing energy function.

For $\rho_\theta(e)$, given diversity in gene variations, predefining distribution presents difficulty. In RRC model, unknown PDF $\rho_\theta(e)$ is assumed symmetric, differentiable, and monotonic. Therefore, $\rho_\theta(e)$ features the following properties: (1) $\rho_\theta(0)$ is global minimal of $\rho_\theta(Z)$; (2) $\rho_\theta(Z) = \rho_\theta(-Z)$; (3) if $|Z_1| < |Z_2|$, then $\rho_\theta(Z_1) < \rho_\theta(Z_2)$. Without loss of generality, we let $\rho_\theta(0) = 0$. Meanwhile, $\rho_\theta(e)$ is allowed to feature a more flexible shape, which adapts to input testing sample y, to make the system more robust to outliers. Then, by Taylor expansion, Equation (18) can be approximated as follows:

$$\alpha^{\wedge} = argmax_{\alpha}\left\{\frac{1}{2}\| W^{1/2}(y - D\alpha)\|_2^2 + \sum_{j=1}^{n} \rho_\sigma(\alpha_j)\right\} \tag{19}$$

where W is a diagonal matrix and can be updated via the following formula:

1. Set the initial value of iteration count $t = 1$.
2. Compute the coding residual:
$e^{(t)} = y - D\alpha^{(t)}$
where $\alpha^{(1)} = \left[\frac{1}{m};\ \frac{1}{m};\ \cdots;\ \frac{1}{m};\right]$ is the initial vector, and m is the mean of all training samples.
3. Estimate weight value of each gene as follows:
$\omega_\theta(e_i^{(t)}) = 1/\left(1 + \exp\left(-\mu(e_i^{(t)})^2 - \mu\delta\right)\right)$
where μ and δ are estimated in each iteration, and δ is associated with residual.
4. Weighted regularized sparse representation coefficient:
$\alpha^* = \text{argmin}_\alpha\left\{\frac{1}{2}\lVert(w^{(t)})^{0.5}(y - D\alpha)\rVert_2^2 + \sum_{j=1}^{n}\rho_\sigma(\alpha_j)\right\}$
where $w^{(t)}$ is the estimated diagonal weight matrix with $w_{i,i}^{(t)} = \omega_\theta(e_i^{(t)},\ \rho_\sigma(\alpha_j) = \lambda\left\lvert\alpha_j\right\rvert^\beta$ and $\beta = 1$.
5. Update the sparse representation coefficients:
If $t = 1$, $\alpha^{(t)} = \alpha^*$;
If $t > 1$, $\alpha^{(t)} = \alpha^{(t-1)} + \upsilon^{(t)}(\alpha^* - \alpha^{(t-1)})$;
where $0 < \upsilon^{(t)} \leq 1$ is a suitable step size that can be search from 1 to 0 by the standard line-search process[43].
6. Reconstruct the test sample by sparse representation coefficient and all metagenes:
$y_{rec}^{(t)} = D\alpha^{(t)}$ and let $t = t + 1$.
7. Go back to Step 4 until condition of convergence $\lVert W^{(t)} - W^{(t-1)}\rVert_2 / \lVert W^{(t-1)}\rVert_2 < \varphi$, where φ is a small positive scalar) is met, or maximal number of iterations is reached.

Table 7. The RRC algorithm.

Input: Training samples $X = [X_1, X_2, \ldots, X_C] \in R^{m\times n}$
Testing samples $y \in R^m$
Output: Label l of y.
1. Initialize D.
We initialize the atoms of D_i as the eigenvectors of X_i.
2. Update coefficient α.
Fix D and solve α_i, $i = 1, 2, \ldots C$, one by one by solving Eq. (9) with the algorithm presented in Table 5.
3. Update dictionary D.
Fix α and update each D_i, $i = 1, 2, \ldots C$, by solving Eq. (12) with the algorithm presented in Table 6.
4. Classify test sample y.
Fix α and D, and solve the sparse representation $\alpha^\wedge$ of y with the algorithm presented in Table 7.
When the algorithm converges, we can classify the test samples as follows:
$identity(y) = \arg\min_i\left\{\left\lVert W_{final}^{1/2}(y - D_i\alpha_i^\wedge)\right\rVert_2 + wg\lVert\alpha^\wedge - m_i\rVert_2^2,\right.$
where W_{final} is the final weight matrix, $\alpha_{\mathrm{i}}^\wedge$ is the final sub- sparse representation coefficient vector associated with class i, and $\alpha^\wedge$ is the final representation coefficient vector.

Table 8. The FDRRC algorithm.

$$W_{i,i} = \omega_\theta(e_{0,i}) = \rho'_\theta(e_{0,i})/e_{0,i} \tag{20}$$

Thus, minimization of RRC focuses on calculating diagonal weight matrix W. As $\rho_\theta(e)$ is symmetric, differentiable, and monotonic, $\omega_\theta(e_i)$ can be assumed as continuous and symmetric while being inversely proportional to e_i. With these considerations, the logistic function which features the same properties is a good choice for $\omega_\theta(e_i)$[41]. Thus, we can obtain the following:

$$\omega_\theta(e_i) = \exp(-\mu e_i^2 + \mu\delta)/(1 + \exp(-\mu e_i^2 + \mu\delta)) \tag{21}$$

where parameters μ and δ represent two positive scalars. Parameter μcontrols decreasing rate from 1 to 0, and δ controls location of demarcation point. With Equations (20) and (21) and $\rho_\theta(0) = 0$, we can formulate Equation (22):

$$\rho_\theta(e_i) = -\frac{1}{2\mu}(\ln(1 + \exp(-\mu e_i^2 + \mu\delta)) - \ln(1 + \exp(\mu\delta))) \tag{22}$$

For $\rho_\sigma(\alpha)$, we can assume that sparse representation coefficient α_i follows a generalized Gaussian distribution as only the representation coefficients associated with training samples from the target class can feature high absolute values. As we do not know beforehand the class of the test sample, a reasonable prior can be that only a small percent of representation coefficients contains significant values. Then, we can used the following equation:

$$f_\sigma(\alpha_j) = \beta exp\left\{-\left(\frac{\lfloor\alpha\rfloor_j}{\sigma_\alpha}\right)^\beta\right\}/(2\sigma_\alpha\Gamma(1/\beta)) \tag{23}$$

where Γ is the gamma function.

After determining distributions $\rho_\theta(e)$ and $\rho_\sigma(\alpha)$, minimized energy function can be used in the iteratively reweighted RRC (IR^3C) algorithm, which was designed by Yang *et al.*, to solve the RRC model efficiently[26]. The RRC (IR3C) algorithm is described in Table 7.

Local center classifier. Equation (3) is the classification function of SRC-based methods that only consider discrimination capability of representation residuals and not the discrimination capability of representation vectors.

Assuming that m_i is the mean sparse representation coefficient vector of class X_i, mean vector m_i can be viewed as the center of class X_i in the transformed space comprising D. Thus, we label m_i as the local center. For classification of tumor, when y originates from class *i*, residual $\|y - D_i\alpha_i{}^\wedge\|_2^2$should be small while $\left\|y - D_j\alpha_j{}^\wedge\right\|_2^2$, $j \neq i$, should be big. In addition, sparse representation coefficient vector $\alpha^\wedge$ should be close to m_i but far from mean vectors of other classes. Considering the above factors, we define the following classifier:

$$e_i = \|y - D_i\alpha_i{}^\wedge\|_2^2 + w\|\alpha^\wedge - m_i\|_2^2 \tag{24}$$

where w is a parameter for balancing contribution of the two terms to classification. Finally, we can obtain the label of y according to the following formula:

$$identity(y) = \mathrm{argmin}_i(e_i) \tag{25}$$

Algorithm of FDRRC. By combining the IR3C algorithm[26] and Fisher discrimination dictionary learning model, we can obtain the algorithm of FDRRC. Table 8 shows the overall procedure of the algorithm.

References

1. Desai, A. N. & Jere, A. Next Generation Sequencing: ready for the clinics? *Clin Genet* **81**, 503–510 (2012).
2. Li, X. A fast and exhaustive method for heterogeneity and epistasis analysis based on multi-objective optimization. *Bioinformatics* (2017).
3. Sirin, U., Erdogdu, U., Polat, F., Tan, M. & Alhajj, R. Effective gene expression data generation framework based on multi-model approach. *Artificial Intelligence in Medicine* **70**, 41 (2016).
4. Gu, C. *et al*. Global network random walk for predicting potential human lncRNA-disease associations. *Sci Rep* **7**, 12442 (2017).
5. Li, W. *et al*. Maxdenominator Reweighted Sparse Representation for Tumor Classification. *Scientific Reports* **7** (2017).
6. Liao, B. *et al*. Learning a weighted meta-sample based parameter free sparse representation classification for microarray data. *PLoS One* **9**, e104314 (2014).
7. Wang, S. L., Sun, L. & Fang, J. Molecular cancer classification using a meta-sample-based regularized robust coding method. *Bmc Bioinformatics* **15**, 1–11 (2014).
8. Liu, J. X., Xu, Y., Zheng, C. H., Kong, H. & Lai, Z. H. RPCA-Based Tumor Classification Using Gene Expression Data. *IEEE/ACM Transactions on Computational Biology & Bioinformatics* **12**, 964–970 (2015).
9. Gui, J., Wang, S. L. & Lei, Y. K. Multi-step dimensionality reduction and semi-supervised graph-based tumor classification using gene expression data. *Artificial Intelligence in Medicine* **50**, 181 (2010).
10. Guyon, I. Erratum: Gene selection for cancer classification using support vector machines. *Machine Learning* **46**, 389–422 (2001).
11. Sharma, A. & Paliwal, K. K. Cancer classification by gradient LDA technique using microarray gene expression data. *Data & Knowledge Engineering* **66**, 338–347 (2008).
12. Nguyen, D. V. & Rocke, D. M. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**, 39–50 (2002).
13. Wang, S. L., Li, X., Zhang, S., Gui, J. & Huang, D. S. Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction. *Computers in Biology & Medicine* **40**, 179 (2010).
14. Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S. & Ma, Y. Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **31**, 210–227 (2008).
15. Ma, A. P. *et al*. Robust face recognition via gradient-based sparse representation. *Journal of Electronic Imaging* **22**, 3018 (2013).
16. Zheng, C. H., Zhang, L., Ng, T. Y., Shiu, S. C. & Huang, D. S. Metasample-based sparse representation for tumor classification. *IEEE/ACM Transactions on Computational Biology & Bioinformatics* **8**, 1273 (2011).
17. Gan, B., Zheng, C. H. & Liu, J. X. Metasample-Based Robust Sparse Representation for Tumor Classification. *Engineering* **05**, 78–83 (2013).
18. Hang, X. & Wu, F. X. Sparse Representation for Classification of Tumors Using Gene Expression Data. *Journal of Biomedicine & Biotechnology* **2009**, 6, https://doi.org/10.1155/2009/403689 (2009).
19. Chen, X. & Huang, L. LRSSLMDA: Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction. *Plos Computational Biology* **13**, e1005912 (2017).
20. Chen, X., Huang, L., Xie, D. & Zhao, Q. EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction. *Cell Death & Disease* **9**, 3 (2018).
21. Jiang, Z., Lin, Z. & Davis, L. S. Label Consistent K-SVD: Learning A Discriminative Dictionary for Recognition. *IEEE Trans Pattern Anal Mach Intell* **35**, 2651–2664 (2013).
22. Mairal, J., Bach, F. & Ponce, J. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **34**, 791 (2012).
23. Yang, M. & Zhang, L. Gabor Feature Based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary. *European Conference on Computer Vision*. 448–461 (2010).
24. Wang, H., Yuan, C., Hu, W. & Sun, C. Supervised class-specific dictionary learning for sparse modeling in action recognition. *Pattern Recognition* **45**, 3902–3911 (2012).
25. Yang, M., Zhang, L., Feng, X. & Zhang, D. Sparse Representation Based Fisher Discrimination Dictionary Learning for Image Classification. *International Journal of Computer Vision* **109**, 209–232 (2014).

26. Yang, M., Zhang, L., Yang, J. & Zhang, D. Regularized Robust Coding for Face Recognition. *IEEE Transactions on Image Processing* **22**, 1753 (2015).
27. Chun-Hou, Z. Metasample-Based Sparse Representation for Tumor Classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8**, 1273–1282 (2011).
28. Golub, T. R. *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537, https://doi.org/10.1126/science.286.5439.531 (1999).
29. 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).
30. 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).
31. Alizadeh, A. A. *et al*. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
32. Singh, D. *et al*. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203–209 (2002).
33. 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, https://doi.org/10.1016/S1535-6108(02)00032-6 (2002).
34. Armstrong, S. A. *et al*. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* **30**, 41–47, http://www.nature.com/ng/journal/v30/n1/suppinfo/ng765_S1.html (2002).
35. Robnik-Šikonja, M. & Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning* **53**, 23–69, https://doi.org/10.1023/a:1025667309714 (2003).
36. You, Z. H. *et al*. PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. *Plos Computational Biology* **13**, e1005455 (2017).
37. Xing, C., Yan, C. C., Xu, Z. & You, Z. H. Long non-coding RNAs and complex diseases: from experimental results to computational models. *Briefings in Bioinformatics* **18**, 558 (2016).
38. Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern Classification (2nd Edition)*. (Wiley 2001).
39. Beck, A. & Teboulle, M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *Siam Journal on Imaging Sciences* **2**, 183–202 (2009).
40. Yang, A. Y., Ganesh, A., Sastry, S. & Sciences, C. Fast L1-Minimization Algorithms and An Application in Robust Face Recognition: A Review. 1849–1852 (2010).
41. Ziegel, E. R. The Elements of Statistical Learning. *Springer* **167**, 192–192 (2003).
42. Wright, S. J., Nowak, R. D. & Figueiredo, M. A. T. Sparse Reconstruction by Separable Approximation. *IEEE Transactions on Signal Processing* **57**, 2479–2493 (2009).
43. Hiriart-Urruty, J. B. & Lemaréchal, C. *Convex Analysis and Minimization Algorithms I*. **1**, 150–159 (2001).

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 Number: 11171369, 61272395, 61370171, 61300128, 61472127, 61572178, 61672214, 61672223 and 61772192), Nature Science Foundation of Hunan Province (Grant No. 2016JJ4029, No. 2018JJ2055).

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., Z.J.L., X.H.W., C.L.G. implemented the experiments, and analyzed the result. H.G.H., L.J.C., H.W.C. analyzed the result. All authors reviewed the final manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

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