On Efficient Feature Ranking Methods for High-Throughput Data Analysis

Bo Liao, Yan Jiang, Wei Liang, Lihong Peng, Li Peng, Damien Hanyurwimfura, Zejun Li, and Min Chen

Abstract—Efficient mining of high-throughput data has become one of the popular themes in the big data era. Existing biology-related feature ranking methods mainly focus on statistical and annotation information. In this study, two efficient feature ranking methods are presented. Multi-target regression and graph embedding are incorporated in an optimization framework, and feature ranking is achieved by introducing structured sparsity norm. Unlike existing methods, the presented methods have two advantages: (1) the feature subset simultaneously account for global margin information as well as locality manifold information. Consequently, both global and locality information are considered. (2) Features are selected by batch rather than individually in the algorithm framework. Thus, the interactions between features are considered and the optimal feature subset can be guaranteed. In addition, this study presents a theoretical justification. Empirical experiments demonstrate the effectiveness and efficiency of the two algorithms in comparison with some state-of-the-art feature ranking methods through a set of real-world gene expression data sets.

Index Terms—Feature ranking, $\ell_{2,1}$-norm, microarray data analysis, convex optimization, regression, manifold learning

1 Introduction

Big data mining methods have gained tremendous attention in the fields of modern pattern recognition, image processing, and bioinformatics. Biomedical researchers studied high-throughput data such as gene expression data [1], protein mass spectrometry data [2], and single nucleotide polymorphism data [3] because the gene expression level in a cell could reflect the risk of cancer [4]. Moreover, only the malignant tumors are called cancer. Consequently, differential gene expression analysis is greatly significant in biomarker discovery and early diagnosis.

Microarray data is a typical high-throughput data. However, only a few samples can be measured experimentally because biology experiment are costly. The small sample size problem poses a great challenge to conventional data mining schemes. Unfortunately, the dimension of microarray data is often higher than the number of samples, and this drawback further increases the difficulty of classification, since the curse of dimensionality and over-fitting are inevitable in such situations [5].

To alleviate the shortcomings mentioned above, dimensionality reduction approaches have been widely studied recently. When a compact subset of informative features is available, traditional data analysis techniques can be applied. Generally, dimensionality reduction methods roughly fall into two categories: feature extraction and feature selection. The former transforms high-dimensional data into low-dimensional data by mapping [6]. However, finding the significance of data with the new representation is difficult. Feature ranking is more appropriate when the significance is emphasized. Feature ranking methods can be categorized into filter method, embedded method and wrapper method. Many filter methods [7], [8], [9], [10], [11] have been proposed for simplicity and computational efficiency. Fisher score [8] ranking feature is based on the optimization criterion to maximize the ratios of between-class and within-class covariance. Laplacian score is proposed by He et al. [9], it's an unsupervised variant of Fisher score and the locality information is depicted by a k-nearest neighbor graph. Statistical ranking based methods are usually applied in differential expression gene analysis [10], [11]. One shortcoming of these methods is the individual selection of features, which ignores the correlation between features. Thus, this greedy strategy-based method often leads to sub-optimal results. Microarray data is high-noise data with small sample size, consequently, the statistical-based methods may be unreliable to some extent. The embedded method such as [12] has been successfully used in gene clustering, but the solution depends heavily on eigen value decomposition in gene elimination processing. The process is time consuming when the number of genes becomes tens of thousands. In [13], the interaction between features is considered. Moreover, the global optimal as a two-stage method cannot be guaranteed, and the feature ranking processing is heuristic and difficult to interpret. Nie et al. [14] selected features in batches by solving trace ratio optimization problem. Unfortunately, the iteration is also based on eigen value decomposition, and it's hard to apply to big data mining. The wrapper method is known for its superior performance as in [15], [16]. However, the computational

- *B. Liao, Y. Jiang, L. Peng, D. Hanyurwimfura, Z. Li, and M. Chen are with the Key Laboratory for Embedded and Network Computing of Hunan Province, the College of Information Science and Engineering, Hunan University, Changsha Hunan 410082, China. E-mail: dragonbw@163.com, jianghnu@hnu.edu.cn.*
- *W. Liang is with the School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China.*
- *L. Peng is with the Department of Computer Science and Engineering, Changsha Medical University, Changsha Hunan 410219, China.*

Manuscript received 11 Aug. 2014; revised 18 Feb. 2015; accepted 26 Feb. 2015. Date of publication 23 Mar. 2015; date of current version 4 Dec. 2015. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TCBB.2015.2415790

cost is the main concern of the wrapper method in addition to the difficulty in extending to multi-class classification problem.

In this study, two novel feature ranking methods, namely, efficient feature ranking via $\ell_{2,1}$-regularization (EFRL21) and robust efficient feature ranking via $\ell_{2,1}$-regularization (REFRL21) are presented. In our methods, the global margin information and locality manifold information are considered for feature ranking. The main contributions of this study can be highlighted as follows:

- The proposed methods not only share the advantages of multi-target regression that global between-class margin structure preserved but also have the desirable characteristics of embedding the discriminating information in locality manifold structure. The structured sparsity norm is added as a regularization term for feature ranking. Recent studies [6], [9] showed that locality information is essential for recognition.
- Compared with existing methods, the proposed methods select differential expression genes by batch rather than individually. Thus, the correlations between genes are considered. Cai et al. [13] also selected features using batch model. However, this two-stage based method cannot guarantee global optimal results. The proposed methods have a general framework, and any kind of high-throughput data can be incorporated within our methods.
- Theoretically, two iterative-based algorithms are presented, and the convergence is analyzed in depth. Gradient information is avoided in each iteration. Thus, our methods converge quickly. Extensive experiments are presented to show the effectiveness of the proposed methods in terms of five evaluation metrics. These promising metiers provide new insights for big data mining methods in bioinformatics.

The rest of this study is organized as follows: Some notations are introduced in Section 2. Formulations of the proposed methods and theoretical justifications are presented in Section 3. Extensive experiments that demonstrate the efficacy of the novel algorithms are reported in Section 4. The proposed methods are discussed in Section 5. The conclusion and suggestions for further research are presented in Section 6.

2 Notations

This section introduces the mathematical notations used in this study. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m] \in \mathbb{R}^{n \times m}$ denotes a data matrix, m and n denote the number of samples and features, respectively. We use $\mathbf{x}_i \in \mathbb{R}^n$ to denote the ith sample in data matrix, it's a column vector with n dimension. $\mathbf{Y} \in \mathbb{R}^{c \times m}$ is the target matrix, if the ith sample belongs to the jth class $y_{ji} = 1$, 0 otherwise. Let $\mathbf{y}_i \in \mathbb{R}^c$ denotes the class label vector with c dimension of the ith sample. For transformation matrix $\mathbf{W} \in \mathbb{R}^{n \times c}$, we use $\mathbf{w}^i$ and $\mathbf{w}_j$ to denote the ith row and jth column of $\mathbf{W}$, respectively. $\ell_{2,1}$-norm, which was proposed by Ding et al. [17] and known for its rotation invariant merits. The definitions of Frobenius norm and $\ell_{2,1}$-norm for matrix are as Eq. (1) and Eq.(2). Matrices used in this study are in boldface uppercase letters, and vectors are marked with boldface lowercase letters,

$$\|\mathbf{W}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{c} w_{ij}^2} = \sqrt{\sum_{i=1}^{n} \|\mathbf{w}^i\|_2^2} \tag{1}$$

$$\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{c} w_{ij}^2} = \sum_{i=1}^{n} \left\|\mathbf{w}^i\right\|_2. \tag{2}$$

3 Efficient Feature Ranking Methods for High-Throughput Data Analysis

In this section, we first briefly review some useful conceptions related to our study, and then we present the mathematical model of EFRL21 and REFRL21. Finally, we discuss the optimization methods as well as the theoretical study about convergence.

3.1 Formulations

Considering multi-target regression problem in Eq. (3), $\mathbf{b}$ is an intercept term in linear systems. The systems that transform $\mathbf{x}_i$ to its target $\mathbf{y}_i$ use linear transformation. The target in Eq. (3) can be regarded as a measurement of margin [18], given that only the elements that corresponding to a certain class are equal to one and the remaining elements are equal to zero, the global margin structure can be established through regression,

$$\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i + \mathbf{b} \tag{3}$$

$$A_{i,j} = \begin{cases} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{t^2}} & \text{if } \mathbf{x}_\mathrm{i} \in \mathcal{L}(\mathbf{x}_\mathrm{j}) \text{ or } \mathbf{x}_\mathrm{j} \in \mathcal{L}(\mathbf{x}_\mathrm{i}). \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Based on manifold assumption [9], [19], high-dimensional data reside on a low-dimensional submanifold embedded in ambient space. Two data points that are proximate and share the same class label should be closer after mapping from the original space. This local structure information has been successfully used in clustering [13]. Locality structure information can be more precisely captured when the class label is available. In this study, we use heat kernel to measure the affinity between samples, as defined in Eq. (4), $\mathcal{L}(\mathbf{x}_j)$ denotes a set of samples that share the same class label with $\mathbf{x}_j$. Specifically, many kernel functions can be used to measure the similarity:

1) *Binary kernel*. $w_{i,j} = 1$ if and only if two samples that come from the same class; otherwise the value is 0. This kernel is simple and computationally efficient.
2) *Cosine kernel*. If sample i and j share the same label, then the value is $w_{i,j} = \frac{\mathbf{x}_i^T\mathbf{x}_j}{\|\mathbf{x}_i\|_2\left\|\mathbf{x}_j\right\|_2}$; 0 otherwise.
3) *Heat kernel*. This was used in this study and defined in Eq. (4),

$$\min_{\mathbf{W}} \frac{1}{2}\sum_{i,j=1}^{m} \left\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right\|^2 A_{ij}. \tag{5}$$

To ensure that the mapped points that share the same label are sufficiently closer, one can minimize Eq. (5) after performing some simple algebraic steps, Eq. (5) can be integrated into Eq. (6), where $\mathbf{L} = \mathbf{D} - \mathbf{A}$ and $\mathbf{L}$ is named as Laplacian matrix. $\mathbf{D}$ is a diagonal matrix, and D_{ii} is the column sum of $A_{i,j}$,

$$\begin{aligned}&\sum_{i,j=1}^{m} \left\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right\|^2 A_{ij} \\ &= \frac{1}{2}\sum_{i,j} Tr\left(\left(\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right)\left(\mathbf{x}_i^T\mathbf{W} - \mathbf{x}_j^T\mathbf{W}\right)\right) A_{ij} \\ &= Tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}).\end{aligned} \tag{6}$$

We can easily verify that in Eq. (5), if two samples that share the same label are mapped too far from each other, then a penalty will be incurred in the objective function. Therefore, the local structure measured by the heat kernel can be well preserved,

$$\begin{aligned}\min_{W,b} &\sum_{i=1}^{m} \left\|\mathbf{W}^T\mathbf{x}_i + \mathbf{b} - \mathbf{y}_i\right\|^2 \\ &+ \frac{\lambda}{2}\sum_{i,j=1}^{m} \left\|\mathbf{W}^T\mathbf{x}_i - \mathbf{W}^T\mathbf{x}_j\right\|^2 A_{ij} + \gamma\|\mathbf{W}\|_{2,1}\end{aligned} \tag{7}$$

$$\begin{aligned}\min_{\mathbf{W},\mathbf{b}} &\left\|\mathbf{W}^T\mathbf{X} + \mathbf{b}\mathbf{1}^T - \mathbf{Y}\right\|_F^2 + \lambda Tr(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) \\ &+ \gamma\|\mathbf{W}\|_{2,1}.\end{aligned} \tag{8}$$

The key idea in our study is to determine a subset of features that respect both global and local structure in data. Intuitively, in our objective function Eqs. (7) and (8) are composed of three parts. The first part considers the global margin structure when least square regression is minimized; the second part considers the local manifold structure, $\ell_{2,1}$-norm is added in our objective function to guarantee that some row coefficients in $\mathbf{W}$ will shrink to zero, which can be regarded as a special case of group lasso [20]. This detail is essential for feature selection because the coefficients' magnitude in regression can measure the importance of a variable, and each row in $\mathbf{W}$ corresponds to a feature. The $\ell_{2,1}$ norm can guarantee row sparsity of transformation matrix $\mathbf{W}$. Thus, the irrelevant features are discarded. Moreover, the proposed methods select features according to a batch model unlike existing schemes. For simplicity, the intercept term can be absorbed into $\hat{\mathbf{W}}$ as in Eq. (9) if we append a new vector $\mathbf{1}^T$ with all its elements equal to 1 in $\hat{\mathbf{X}}$ (see Eq. (10)), and the objective function of EFRL21 can be reformulated as Eq. (11),

$$\hat{\mathbf{W}} = \begin{bmatrix}\mathbf{W} \\ \mathbf{b}^T\end{bmatrix} \in \mathbb{R}^{(n+1)\times c} \tag{9}$$

$$\hat{\mathbf{X}} = \begin{bmatrix}\mathbf{X} \\ \mathbf{1}^T\end{bmatrix} \in \mathbb{R}^{(n+1)\times m} \tag{10}$$

$$\min_{\hat{\mathbf{W}}} \left\|\hat{\mathbf{W}}^T\hat{\mathbf{X}} - \mathbf{Y}\right\|_F^2 + \lambda Tr(\hat{\mathbf{W}}^T\hat{\mathbf{X}}\mathbf{L}\hat{\mathbf{X}}^T\hat{\mathbf{W}}) + \gamma\left\|\hat{\mathbf{W}}\right\|_{2,1} \tag{11}$$

$$\min_{\hat{\mathbf{W}}} \left\|(\hat{\mathbf{W}}^T\hat{\mathbf{X}} - \mathbf{Y})^T\right\|_{2,1} + \lambda Tr(\hat{\mathbf{W}}^T\hat{\mathbf{X}}\mathbf{L}\hat{\mathbf{X}}^T\hat{\mathbf{W}}) + \gamma\left\|\hat{\mathbf{W}}\right\|_{2,1}. \tag{12}$$

Kong et al. [21] indicated that $\ell_{2,1}$-norm is robust to noise and outliers. The extension of EFRL21, which is named as REFRL21 is formulated in Eq. (12). The significant difference between EFRL21 and REFRL21 is the definition of loss function in the Eq. (11) and Eq. (12). F-norm is sensitive to outliers. However, $\ell_{2,1}$-norm is difficult to optimize. In Eq. (12) the residual error is measured in a column-wise manner. Thus, the matrix transport operator is applied in the first term.

3.2 Optimization Method

Notably, objection functions EFRL21 and REFRL21 both are convex optimization problem [22], and the global optimal value can be achieved theoretically. However, the first-order derivative is discontinuous when $\hat{\mathbf{w}} = 0$. Therefore, a closed-form solution cannot be derived directly. Inspired by [23], an iterative based optimization framework is proposed in this study. First, an auxiliary function is introduced in Eq. (15), $\mathbf{U}$ is defined as Eq. (13). We iteratively optimize the auxiliary function. In the following section, we prove that the objective function Eq. (11) would monotonously decrease in each iteration,

$$\mathbf{U} = \begin{pmatrix} 2\left\|\hat{\mathbf{w}}^1\right\|_2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{2\left\|\hat{\mathbf{w}}^{n+1}\right\|_2} \end{pmatrix} \tag{13}$$

$$\left\|\hat{\mathbf{W}}\right\|_{2,1} = 2Tr(\hat{\mathbf{W}}^T\mathbf{U}\hat{\mathbf{W}}) \tag{14}$$

$$\mathcal{J}(\hat{\mathbf{W}}, \mathbf{U}) = \left\|\hat{\mathbf{W}}^T\hat{\mathbf{X}} - \mathbf{Y}\right\|_F^2 + \lambda Tr(\hat{\mathbf{W}}^T\hat{\mathbf{X}}\mathbf{L}\hat{\mathbf{X}}^T\hat{\mathbf{W}}) + \gamma Tr(\hat{\mathbf{W}}^T\mathbf{U}\hat{\mathbf{W}}) \tag{15}$$

$$\frac{\partial\mathcal{J}}{\partial\hat{\mathbf{W}}} = 2\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{W}} - \mathbf{Y}^T) + 2\lambda\hat{\mathbf{X}}\mathbf{L}\hat{\mathbf{X}}^T\hat{\mathbf{W}} + 2\gamma\mathbf{U}\hat{\mathbf{W}} = \mathbf{0} \tag{16}$$

$$\hat{\mathbf{W}} = (\hat{\mathbf{X}}\hat{\mathbf{X}}^T + \lambda\hat{\mathbf{X}}\mathbf{L}\hat{\mathbf{X}}^T + \gamma\mathbf{U})^{-1}\hat{\mathbf{X}}\mathbf{Y}^T. \tag{17}$$

To optimize auxiliary function Eq. (15) when $\mathbf{U}$ is fixed, one can take the derivative of $\mathcal{J}(\hat{\mathbf{W}}, \mathbf{U})$ with respect to $\hat{\mathbf{W}}$ and requires it to be zero, that is, in Eq. (16) and Eq. (17). Then, matrix $\mathbf{U}$ can be updated by Eq. (13).

The procedure EFRL21 is summarized in Algorithm 1.[1]

Similarly, to optimize Eq. (12) the auxiliary function is defined as Eq. (18), where $\mathbf{V}$ is a diagonal matrix and $v_{ii} = \frac{1}{2\left\|\hat{\mathbf{W}}^T\hat{\mathbf{x}}_i - \mathbf{y}_i\right\|_2}$,

$$\begin{aligned}\mathcal{J}(\hat{\mathbf{W}}, \mathbf{V}, \mathbf{U}) = &Tr((\hat{\mathbf{W}}^T\hat{\mathbf{X}} - \mathbf{Y})\mathbf{V}(\hat{\mathbf{W}}^T\hat{\mathbf{X}} - \mathbf{Y})^T) \\ &+ \lambda Tr(\hat{\mathbf{W}}^T\hat{\mathbf{X}}\mathbf{L}\hat{\mathbf{X}}^T\hat{\mathbf{W}}) + \gamma Tr(\hat{\mathbf{W}}^T\mathbf{U}\hat{\mathbf{W}}).\end{aligned} \tag{18}$$

1. To guarantee the convergence of Algorithm 1 a small regularizer term ϵ is added to $u_{ii} = \frac{1}{2\left\|\mathbf{w}^i\right\|_2 + \epsilon}$ when $\|\mathbf{w}^i\|_2 = 0$.

Requiring the derivative of $\mathcal{J}(\hat{\mathbf{W}}, \mathbf{V}, \mathbf{U})$ with respect to $\hat{\mathbf{W}}$ vanish, i.e., $\frac{\partial J(\hat{\mathbf{W}},\mathbf{V},\mathbf{U})}{\partial \hat{\mathbf{W}}} = 0$. We have Eq. (19)

$$\hat{\mathbf{X}}\mathbf{V}(\hat{\mathbf{W}}^T\hat{\mathbf{X}} - \mathbf{Y})^T + \lambda\hat{\mathbf{X}}\mathbf{L}\hat{\mathbf{X}}^T\hat{\mathbf{W}} + \gamma\mathbf{U}\hat{\mathbf{W}} = 0 \tag{19}$$

or equivalently

$$\hat{\mathbf{W}} = (\hat{\mathbf{X}}\mathbf{V}\hat{\mathbf{X}}^T + \lambda\hat{\mathbf{X}}\mathbf{L}\hat{\mathbf{X}}^T + \gamma\mathbf{U})^{-1}\hat{\mathbf{X}}\mathbf{V}\mathbf{Y}^T. \tag{20}$$

The procedure of REFRL21 is summarized in Algorithm 2.[2] Compared with Algorithm 1, Algorithm 2 begins the iteration with matrix $\hat{\mathbf{W}}$ such that all elements are equal to one. By contrast, Algorithm 1 begins the iteration with an identity matrix.

Algorithm 1. Efficient feature ranking via $\ell_{2,1}$-regularization

Input: $\mathbf{U}_0 = \mathbf{I}_{(n+1)\times(n+1)}$;
Input: $t = 0$, $\mathbf{X}$, λ, γ;
Output: feature ranking list;
1: **repeat**
2: Compute $\hat{\mathbf{W}}_{t+1}$ based on Eq. (17);
3: Compute $\mathbf{U}_{t+1}$ based on $\hat{\mathbf{W}}_{t+1}$;
4: $t = t + 1$;
5: **until** convergence
6: **return** feature ranking list based on sorting $\{\|w^i\|_2\}_{i=1}^n$ in descending order;

Algorithm 2. Robust and efficient feature ranking via $\ell_{2,1}$-regularization

Input: $\hat{\mathbf{W}}_0 = \mathbf{1}_{(n+1)\times c}$;
Input: $t = 0$, $\mathbf{X}$, λ, γ;
Output: feature ranking list;
1: **repeat**
2: Compute $\mathbf{V}_{t+1}$ based on $\hat{\mathbf{W}}_t$;
3: Compute $\mathbf{U}_{t+1}$ based on $\hat{\mathbf{W}}_t$;
4: Compute $\hat{\mathbf{W}}_{t+1}$ based on Eq. (20);
5: $t = t + 1$;
6: **until** convergence
7: **return** feature ranking list based on sorting $\{\|w^i\|_2\}_{i=1}^n$ in descending order;

Many approaches have been presented to mathematically solve the $\ell_{2,1}$-norm based regularization problem. However, most of them are based on interior point method or sub-gradient descent scheme, and their convergence is slow. Finding an appropriate step size is time consuming. In this study, we solve the complex mathematical problem by using a simple iterative-based scheme, and we avoided gradient information. Experiential results show that in most cases, the proposed methods converge quickly and the number of iterations are often less than 30.

2. To guarantee the convergence of Algorithm 2 a small regularizer term ϵ is added to $u_{ii} = \frac{1}{2\|\mathbf{w}^i\|_2+\epsilon}$ and $v_{ii} = \frac{1}{2\|\hat{\mathbf{W}}^T\hat{\mathbf{x}}_i - \mathbf{y}_i\|_2+\epsilon}$, respectively. When $\|\mathbf{w}^i\|_2 = 0$ or $\|\hat{\mathbf{W}}^T\hat{\mathbf{x}}_i - \mathbf{y}_i\|_2 = 0$.

3.3 Convergence Analysis

As mentioned in the previous section, EFRL21 and REFRL21 are used for convex optimization and have lower bounds. Hence, based on Cauchy's convergence criterion, we only need to prove the monotonicity of objective functions.

To prove the monotonicity, we need to prove the lemma as follows:

Lemma 1. *For any non-zero real-number vectors* $\mathbf{a}$, $\mathbf{b}$ *we have the following inequality:*

$$\frac{\|\mathbf{a}\|_2^2}{2\|\mathbf{b}\|_2} - \|\mathbf{a}\|_2 \geqslant \frac{\|\mathbf{b}\|_2^2}{2\|\mathbf{b}\|_2} - \|\mathbf{b}\|_2. \tag{21}$$

Proof. It's easy to check that

$$\|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 \geqslant 2\|\mathbf{a}\|_2\|\mathbf{b}\|_2 \tag{22}$$

$$\frac{\|\mathbf{a}\|_2^2}{2\|\mathbf{b}\|_2} + \frac{\|\mathbf{b}\|_2}{2} \geqslant \|\mathbf{a}\|_2 \tag{23}$$

$$\frac{\|\mathbf{a}\|_2^2}{2\|\mathbf{b}\|_2} - \|\mathbf{a}\|_2 \geqslant \frac{\|\mathbf{b}\|_2^2}{2\|\mathbf{b}\|_2} - \|\mathbf{b}\|_2. \tag{24}$$

This completes the proof. □

The convergence of Algorithm 1 is summarized as in the following theorem:

Theorem 1. *The objective (11) will monotonically decrease in each iteration when we apply algorithm 1 to optimize the problem (15).*

Proof. Suppose that in the tth iteration we have

$$\begin{aligned}\hat{\mathbf{W}}_{t+1} = \arg\min_{\hat{\mathbf{W}}} \mathcal{J}(\hat{\mathbf{W}}, \mathbf{U}_t) = \arg\min_{\hat{\mathbf{W}}} \left\|\hat{\mathbf{W}}^T\hat{\mathbf{X}} - \mathbf{Y}\right\|_F^2 \\ + \lambda Tr(\hat{\mathbf{W}}^T\hat{\mathbf{X}}\mathbf{L}\hat{\mathbf{X}}^T\hat{\mathbf{W}}) + \gamma Tr(\hat{\mathbf{W}}^T\mathbf{U}_t\hat{\mathbf{W}}).\end{aligned} \tag{25}$$

Let

$$\mathcal{Q}(\hat{\mathbf{W}}_{t+1}) = \left\|\hat{\mathbf{W}}_{t+1}^T\hat{\mathbf{X}} - \mathbf{Y}\right\|_F^2 + \lambda Tr\left(\hat{\mathbf{W}}_{t+1}^T\hat{\mathbf{X}}\mathbf{L}\hat{\mathbf{X}}^T\hat{\mathbf{W}}_{t+1}\right).$$

Which indicates that

$$\begin{aligned}\mathcal{Q}(\hat{\mathbf{W}}_{t+1}) + \gamma Tr\left(\hat{\mathbf{W}}_{t+1}^T\mathbf{U}_t\hat{\mathbf{W}}_{t+1}\right) \leqslant \mathcal{Q}(\hat{\mathbf{W}}_t) \\ + \gamma Tr\left(\hat{\mathbf{W}}_t^T\mathbf{U}_t\hat{\mathbf{W}}_t\right),\end{aligned} \tag{26}$$

where $\left\|\mathbf{w}_t^i\right\|_2$ denotes the 2-norm of the ith row in $\hat{\mathbf{W}}$ in the tth iteration,

$$\mathcal{Q}(\hat{\mathbf{W}}_{t+1}) + \gamma\sum_{i=1}^{n+1}\frac{\left\|\mathbf{w}_{t+1}^i\right\|_2^2}{2\|\mathbf{w}_t^i\|_2} \leqslant \mathcal{Q}(\hat{\mathbf{W}}_t) + \gamma\sum_{i=1}^{n+1}\frac{\left\|\mathbf{w}_t^i\right\|_2^2}{2\|\mathbf{w}_t^i\|_2}. \tag{27}$$

It's easy to check that

$$\begin{aligned}\mathcal{Q}(\hat{\mathbf{W}}_{t+1}) + \gamma\sum_{i=1}^{n+1}\left\|\mathbf{w}_{t+1}^i\right\|_2 + \gamma\sum_{i=1}^{n+1}\left(\frac{\left\|\mathbf{w}_{t+1}^i\right\|_2^2}{2\|\mathbf{w}_t^i\|_2} - \left\|\mathbf{w}_{t+1}^i\right\|_2\right) \\ \leqslant \mathcal{Q}(\hat{\mathbf{W}}_t) + \gamma\sum_{i=1}^{n+1}\left\|\mathbf{w}_t^i\right\|_2 + \gamma\sum_{i=1}^{n+1}\left(\frac{\left\|\mathbf{w}_t^i\right\|_2^2}{2\|\mathbf{w}_t^i\|_2} - \left\|\mathbf{w}_t^i\right\|_2\right).\end{aligned} \tag{28}$$

Recalling Lemma 1, we can conclude that

$$\mathcal{Q}(\hat{\mathbf{W}}_{t+1}) + \gamma \sum_{i=1}^{n+1} \left\|\mathbf{w}_{t+1}^{i}\right\|_2 \leqslant \mathcal{Q}(\hat{\mathbf{W}}_t) + \gamma \sum_{i=1}^{n+1} \left\|\mathbf{w}_{t}^{i}\right\|_2. \quad (29)$$

This completes the proof of monotonicity. □

Similarly, the convergence of Algorithm 2 is summarized as in the following theorem:

Theorem 2. *The objective (12) will monotonically decrease in each iteration when we apply algorithm 2 to the optimization problem (18).*

Proof. Please refer to supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2015.2415790 available online. □

3.4 Computational Complexity Analysis

In this section, we analyze the time complexity of Algorithms 1 and 2.

The most expensive computation is the matrix inversion problem in the iteration processing, namely, line 2 and line 4 in EFRL21 and REFRL21, respectively. A matrix can be inversed within $\mathcal{O}(n^3)$. When n becomes large, the computational burden is heavy. Fortunately, one can use Woodbury matrix identity [24] to identify and simplify the matrix inversion problem. For arbitrary matrices $\mathbf{U}$ and $\mathbf{V}$ and invertible matrices $\mathbf{A}$, $\mathbf{C}$, Eq. (30) holds. Inversion of a diagonal matrix is simple, and inversion of the second term in the Woodbury formula can be achieved within $\mathcal{O}(m^3)$. For microarray data, m is very small (less than 100), and the computational processing can be significantly accelerated. The total time complexity cannot be very precisely approximated by $\mathcal{O}(\#iteration \times m^3)$, where $\#iteration$ is the number of iterations,

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}. \quad (30)$$

4 Experiments

In this section, we conducted extensive experiments to evaluate our methods. First, we briefly describe the datasets and compare the algorithms used in our experiments. Then, some impressive experimental results are presented. Finally, we further discuss the property of our methods, such as through convergence analysis.

4.1 Data Sets Descriptions and Compared Algorithms

Six gene expression profile data sets are used in our experiment. The statistics of these data sets are summarized in Table 1. Central Nervous System (CNS), Colon, DLBCL, Leukemia, SRBCT, and MLLLeukemia are publicly available at the Kent Ridge Bio-medical Repository [25].

- Central Nervous System [26] contains 60 patient samples with 7,129 genes in each sample. The class label has two states, 21 survivors are labeled as "Class1" and 39 failures are labeled as "Class0".
- Colon [27] contains 62 samples with 2,000 genes in each sample. The data set is collected from colon-cancer patients. The class label has two states, 40 tumor biopsies are labeled as "negative" and 22 normal samples are labeled as "positive".
- DLBCL [28] contains 77 samples with 7,129 genes in each sample. The class label has two states, 58 diffuse large b-cell lymphoma samples and 19 follicular lymphoma samples are contained in this data set.
- Leukemia [1] contains 72 samples with 7,129 genes in each sample. The class label has two states, 47 acute lymphoblastic leukemia (ALL) samples and 28 acute myeloid leukemia (AML) samples are contained in this data set.
- SRBCT [29] contains 83 samples with 2,308 genes in each sample. The class label has four states, 29 Ewing's sarcoma (EWS) samples, 11 Burkitt's lymphoma (BL) samples, 18 neuroblastoma (NB) samples and 23 rhabdomyosarcoma (RMS) samples are contained in this data set.
- MLLLeukemia [4] contains 72 samples with 12,582 genes in each sample. The class label has three states, 24 acute lymphoblastic leukemia samples, 28 acute myeloid leukemia samples and 20 mixed lineage leukemia protein (MLL) samples are contained in this data set.

TABLE 1
Summary of the Data Sets Used in Our Experiments

Datasets	Samples	Subclass number	Genes
Central Nervous System	60	2	7,129
Colon	62	2	2,000
DLBCL	77	2	7,129
Leukemia	72	2	7,129
SRBCT	83	4	2,308
MLLLeukemia	72	3	12,582

To illustrate how recognition performance can be improved by our methods, we compare the following five supervised feature ranking schemes:

- Kruskal Wallis one way analysis of variance (KW) [30] is a novel statistical ranking (variance information) based feature selection method. It has been widely used in differential gene expression analysis [10], [11] for the merits that KW is a non-parametric method and does not relay on data distribution.
- Locality sensitive Laplacian score (LSLS) [7] is a variant of spectral feature selection approach, the differential expression genes are selected based on their contributions to the local margin preserve capability and variance information. The local margin structure and variance information are assessed by spectral graph theory [31] in LSLS.
- Robust feature selection (RFS) [23] formulates margin structure as a multi-target regression optimization problem, and informative genes are selected by structured sparsity norm.
- Efficient feature ranking via $\ell_{2,1}$-regularization (EFRL21), compared with RFS, EFRL21 takes manifold structure into consideration as a regularizer

term. Consequently, the global information and local manifold information are considered simultaneously in our optimization framework.

- Robust and efficient feature ranking via $\ell_{2,1}$-regularization (REFRL21) is compared with EFRL21. The only difference between them is that we use $\ell_{2,1}$-norm to measure the regression residues rather than F-norm in the first term of our objective function. Motivated by Kong et al. [21] $\ell_{2,1}$-norm is robust to noise and outliers. We use $\ell_{2,1}$-norm to address the high-noise problem in microarray data analysis.

It's worth noting that these five algorithms belong to three families, namely, statistical test (KW), spectral feature selection (LSLS) and structured sparsity learning (RFS, EFRL21, REFRL21). LSLS selects differential expression gene based on local information, RFS selects differential expression gene based on global information, and EFRL21 and REFRL21 consider local and global information simultaneously. In the following section, we comprehensively evaluate the effectiveness of the proposed feature ranking methods.

4.2 Experimental Setup and Evaluation Metrics

In our approaches, three parameters need to be initialized: two regularizer parameters λ and γ control the importance of manifold regularization term and sparsity term, respectively. Heat kernel bandwidth t controls the affinity matrix defined in Eq. (4). In our experiments, we simply set $t = 1$, $\lambda = 0.1$, and $\gamma = 0.001$ for all datasets. For LSLS, we initialize the parameters as [7] suggested and choose the best parameters in RFS by cross-invalidation on training data. To obtain a fair evaluation, we apply stratified sampling technique for each data set, consequently, we use 60 percent of the samples as the training set and the rest for evaluation. The classification results averaged over random 10 test/training splits,

- $$accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$
- $$precision = \frac{TP}{TP + FP}$$
- $$recall = \frac{TP}{TP + FN}$$
- $$fscore = \frac{2 \times precision \times recall}{precision + recall}.$$
- The area under receive operating characteristic curve (AUC) [32] is introduced to quantify the results.

Since the tumor classification problem is a cost-sensitive learning problem [33]. Therefore, the cost is different when we wrongly classify a normal-sample into a case sample, and vice versa. Inspired by Piao et al. [34] we use five metrics, namely, accuracy, precision, recall, fscore and AUC, to perform a comprehensive evaluation. The five metrics range from the interval 0 to 1 and the larger number is the ideal. TP and TN denote the ratio of samples that are correctly classified as positive and negative classes, respectively; FN and FP denote the ratio of samples that are wrongly classified from positive samples to negative samples and negative to positive samples, respectively. The positive and negative sets are defined the same as [34]. All features are normalized with zero mean and unit variance in our pre-processing steps. Finally, LIBSVM [35] with linear kernel is applied on the selected genes for classification. Note that, SRBCT and MLLLeukemia are two multi-class classification problems in which case the one-against-rest method is conducted for each subclass. Thus, we report the averaged results of c binary classification problems, where c denotes the number of subclasses. The number of genes that should be chosen for further investigation is an open question. In this study, we follow literature [36] as Wang indicates that the top 10 percent genes may be an appropriate choice. Thus, we selected the top 400 genes for comprehensive performance evaluation. In the following section, our experiments are implemented in MATLAB environment and run on a personal computer.

4.3 Comprehensive Performance Evaluation

To ensure a fair evaluation, the selected top 50 and 100 genes are used for illustration, as shown in Figs. 1 and 2, which show the classification results on six gene expression data sets, respectively. When 50 genes are used for evaluation, the proposed methods perform well on CNS and COLON data sets, in Figs. 1a and 1b, REFRL21 outperforms its competitors regardless of the metrics that were used, EFRL21, RFS, and LSLS obtain the second best results. In general, REFRL21 demonstrates greater accuracy compared with other methods, the accuracy of REFRL21 improves by at least 5 percent. LSLS, RFS, EFRL21, and REFRL21 exhibit similar performance on SRBCT. KW shows the worst performance. LSLS performs better on Leukemia as shown in Fig. 1d. On two multi-class data sets SRBCT in Fig. 1e and MLLLeukemia in Fig. 1f, EFRL21 and REFRL21 exhibit similar performance in terms of all metrics and show promising results. In sum, REFRL21 and EFRL21 perform better, followed by RFS then LSLS, and lastly by KW.

When more genes such as the top 100 genes are used for evaluation, the performance of all approaches improves generally, see Fig. 2. On CNS, COLON and SRBCT, EFRL21 and REFRL21 achieve the highest classification results, see Figs. 2a, 2b, and 2c. On CNS, KW exhibits great improvement. However, this is an extraordinary case, see Fig. 2a. In most cases, one can observe the same performance trends as in Fig. 1.

Tables 2, 3, and 4 summarize the best classification result that the five schemes can achieve in terms of the six data sets we used. Intuitively, for the small sample size problem and the class imbalance problem, accuracy may be insufficient for evaluation. Therefore, fscore and AUC are computed as well, and the best results are shown in bold letters. Table 2 shows the detailed classification accuracies and the number of selected genes are reported. A comparison with the best algorithm other than the proposed EFRL21 and REFRL21, that is, KW and LSLS, shows that REFRL21 can achieve 5 percent improvement on CNS and COLON. In other cases such as Leukemia, EFRL21 achieves the same accuracy using few genes. In Table 3, EFRL21 and REFRL21 achieve 100 percent on MLLLeukemia and achieve 5 percent

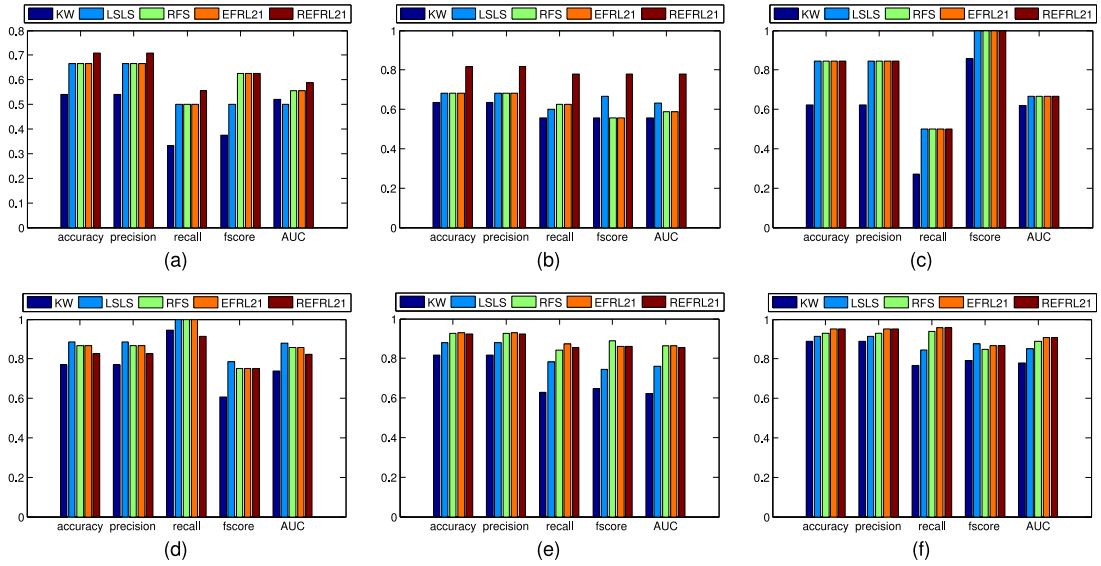


Fig. 1. Classification results with respect to different evaluation metrics on six microarray datasets while the top 50 genes are selected for evaluation. (a) CNS, (b) Colon, (c) DLBCL, (d) Leukemia, (e) SRBCT, and (f) MLLLeukemia.

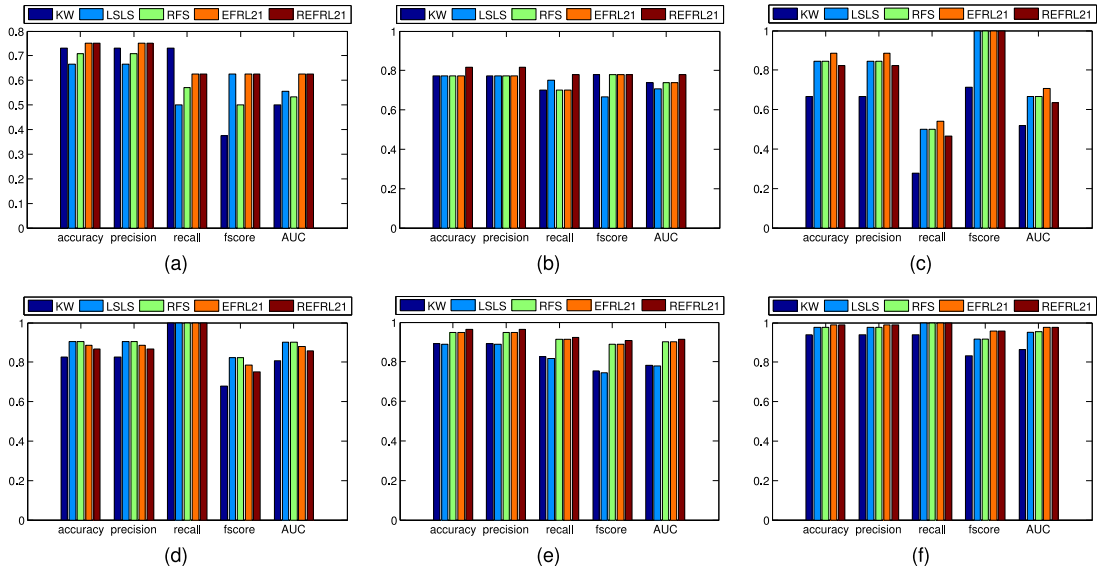


Fig. 2. Classification results with respect to different evaluation metrics on six microarray datasets while the top 100 genes are selected for evaluation. (a) CNS, (b) Colon, (c) DLBCL, (d) Leukemia, (e) SRBCT, and (f) MLLLeukemia.

TABLE 2
Comparison of Five Gene Selection Approaches' Best Prediction Accuracy (%) as Well as the Number of Selected Genes Are Presented in Parentheses

Microarry Datasets	KW(#gene)	LSLS(#gene)	RFS(#gene)	EFRL21(#gene)	REFRL21(#gene)
CNS	75.00(90)	75.00(114)	79.17(36)	75.00(28)	**79.56**(21)
Colon	77.27(13)	81.82(21)	81.82(65)	81.82(65)	**86.36**(11)
DLBCL	84.44(296)	**93.33**(225)	93.33(374)	93.33(274)	93.33(311)
Leukemia	88.46(22)	90.38(57)	94.23(366)	**94.23**(340)	94.23(388)
SRBCT	92.65(357)	94.85(222)	**96.32**(141)	96.32(294)	96.32(300)
MLLLeukemia	94.75(95)	95.75(100)	97.85(77)	98.63(90)	**98.63**(80)

improvement. REFRL21 achieves 88.89 percent using only 12 genes on Colon and achieve 10 percent improvement. AUC is an essential metric for imbalanced data classification. In Table 4, REFRL21 can achieve 6 percent improvement over RFS on CNS and 7 percent on Colon. In other data sets, the proposed methods can achieve comparable results with the competitors. These findings confirm the effectiveness of the proposed schemes.

TABLE 3
Comparison of Five Gene Selection Approaches' Best F-Score (%) as Well as the Number of Selected Genes Are Presented in Parentheses

Microarry Datasets	KW(#gene)	LSLS(#gene)	RFS(#gene)	EFRL21(#gene)	REFRL21(#gene)
CNS	62.5(295)	**75**(19)	75(68)	**75**(19)	**75**(19)
Colon	77.78(15)	77.78(17)	77.78(56)	77.78(54)	**88.89**(12)
DLBCL	85.47(17)	100(39)	100(13)	**100**(12)	100(13)
Leukemia	78.57(21)	82.14(45)	**89.29**(300)	89.29(342)	89.29(392)
SRBCT	84.39(320)	85.01(187)	**90.50**(141)	90.50(175)	90.50(301)
MLLLeukemia	95.5(106)	95.83(14)	93.56(32)	**100**(33)	**100**(33)

TABLE 4
Comparison of Five Gene Selection Approaches' Best AUC Score (%) as Well as the Number of Selected Genes Are Presented in Parentheses

Microarry Datasets	KW(#gene)	LSLS(#gene)	RFS(#gene)	EFRL21(#gene)	REFRL21(#gene)
CNS	62.50(306)	63.16(77)	70.59(33)	63.16(24)	**76.67**(28)
Colon	73.68(107)	77.78(26)	77.78(73)	77.78(64)	**84.21**(15)
DLBCL	66.67(305)	**82.35**(233)	82.35(385)	82.35(284)	82.35(328)
Leukemia	88.00(25)	90.20(53)	**94.34**(305)	94.34(305)	94.34(392)
SRBCT	84.85(362)	89.38(221)	92.83(305)	92.83(291)	**93.78**(304)
MLLLeukemia	92.89(13)	97.73(17)	96.08(324)	**98.08**(332)	98.08(336)

4.4 Convergence Study

Computational complexity is the main concern of our two approaches. In the previous section we have analyzed the time complexity theoretically. However, we have yet to know the time expenditure, since the number of iterations ($\#iteration$) depends heavily on the data set that was used. Here, we investigate the convergence rate of the iterative rules. We present only the convergence curve of EFRL21 because REFRL21 has the similar results.

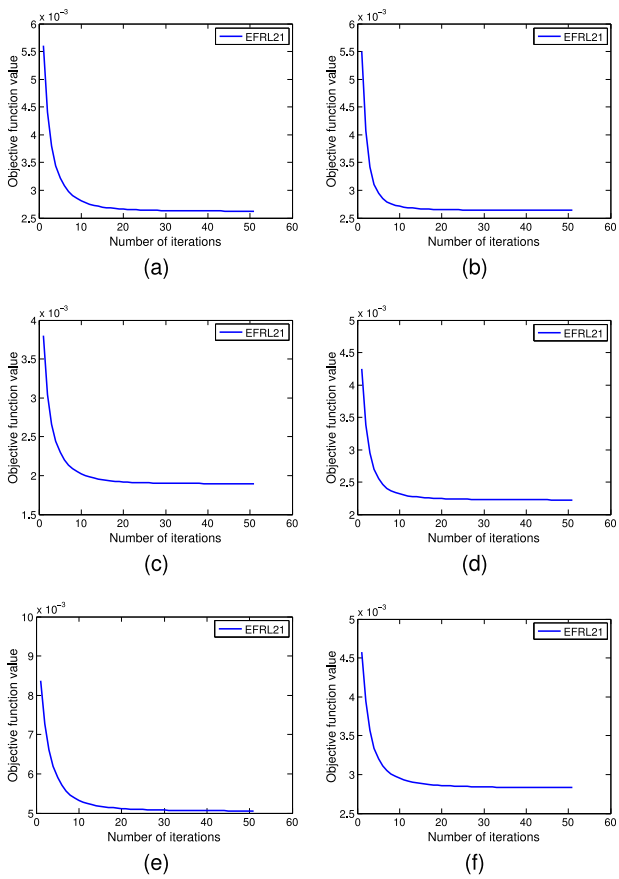


Fig. 3. Convergence curve of EFRL21. (a) CNS, (b) Colon, (c) DLBCL, (d) Leukemia, (e) SRBCT, and (f) MLLLeukemia.

Fig. 3 shows the convergence curves of EFRL21 on six microarray data sets. The y-axis denotes the objective function value, and the x-axis denotes the number of iterations. Obviously, EFRL21 converges quickly. In most cases, the proposed method converges within 30 iterations.

4.5 Parameters Selection Problem

In this subsection, we investigate the influences of two parameters on our objective function when parameters λ and γ are varied from $\{0.001, 0.01, 0.1, 1, 10, 100, 1{,}000\}$ using grid searching. The parameter λ controls the penalty term of manifold regularizer. When λ is larger the samples that share the same label should be closer after mapping to minimize the objective function. Parameter γ controls the row sparsity of $\hat{\mathbf{W}}$ and plays an important role in gene selection. For supervised learning methods, the model selection problem can be easily addressed by performing cross-validation. The classification accuracy of EFRL21 (REFRL21 has the same results and is therefore not discussed) w.r.t. λ and γ are shown in Fig. 4.

As it can be seen from Figs 4a, 4c, 4d and 4e that EFRL21 is stable with respect to different parameters when classification accuracy is taken as a function of the two parameters. Moreover, EFRL21 remains stable over a large range, see Figs. 4b and 4f. The performance exhibits litter fluctuation only when λ becomes extremely large and γ becomes extremely small. From these results, we can conclude that the classification results are insensitive in terms of parameters. Parameters optimization is not very crucial. The model selection part can be avoided because of this superior

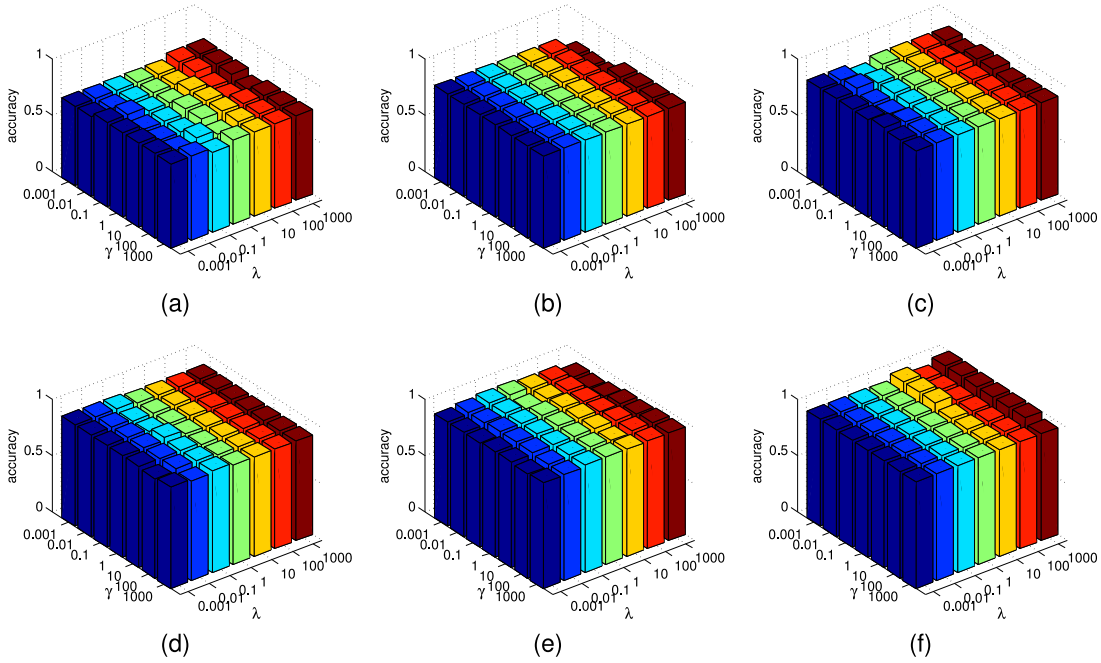


Fig. 4. Classification accuracy with respect to different parameters γ and λ on microarray datasets while the top 150 genes are selected for evaluation. (a) CNS, (b) Colon, (c) DLBCL, (d) Leukemia, (e) SRBCT, and (f) MLLLeukemia.

property, which can further enhance the efficiency of the proposed schemes and provide new insight into biological big data mining methods.

4.6 Effectiveness of Weighting Scheme

The manifold structure is captured by heat kernel function in all our experiments. To further investigate the effectiveness of the weighting scheme, we compare the performance of EFRL21 and REFRL21 with respect to different weighting schemes that were mentioned in the previous section.

In the evaluation processing, 100 top-ranked genes are selected by EFRL21 and REFRL21 on CNS data set. Fig. 5 shows the classification accuracy versus the number of selected genes in terms of three weighting strategies. Two conclusions can be drawn from Fig. 5: (1) either EFRL21 or REFRL21 incorporated with the heat kernel function outperforms the other two weighting schemes, followed by cosine kernel and binary kernel. Since that heat kernel and cosine kernel can depict data manifold accurately, it's difficult to differentiate locality structure in data by binary kernel. However, the binary kernel is computationally efficient for its simplicity. (2) Compared with EFRL21, the classification tendency of REFRL21 is smoother because of the robust measurement of regression residuals in REFRL21. Consequently, noisy genes are eliminated, which indicates the suitability of REFRL21 for gene expression data sets.

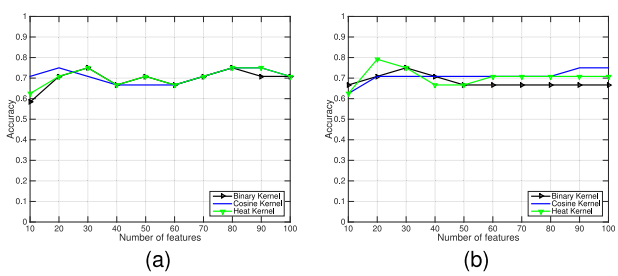


Fig. 5. Performance comparison with respect to different weighting schemes on CNS data set. (a) EFRL21. (b) REFRL21.

5 Discussions

In this section, we discuss the results of the proposed two feature ranking methods and highlight the contributions of this study.

In previous studies [10], [11], many statistical-based methods were employed for microarry data analysis. However, the greatest shortcoming of these methods is that microarry data usually come with very small sample size, which makes the statistical results unreliable. Moreover, gene expression data are often contaminated by noise during quantization processing. Consequently, the performance of KW is the worst in all data sets.

Unlike LSLS, the advantage of EFRL21 and REFRL21 is their insensitivity to the selected parameters. Unfortunately, the performance of LSLS depends heavily on the measurement of locality and parameter optimization, which increases the computational complexity.

RFS considers only global information when solving the multi-target regression. Many studies [7], [37], [38] have shown that locality information is discriminative, and experimental results showed that the performance improved when locality information is considered, see Figs. 1 and 2.

Generally, REFRL21 outperforms EFRL21 in most cases, as shown in Figs. 1a, 1b and Tables 2, 4. This result may be due to the fact that the first term in REFRL21 is measured by $\ell_{2,1}$-norm; the influence of outlier samples and noise features can be alleviated. However, the solution of Algorithm 2 is more complex than 1 even when the two schemes are solved iteratively.

Notably, the proposed methods have a general optimization framework. High-throughput data such as: microarray

data, protein mass spectra data, and Single-nucleotide polymorphism data can all be incorporated in the proposed schemes, which makes the proposed methods scalable.

6 Conclusion and Further Research

In this study, we presented two novel supervised feature ranking methods and their efficient solution framework. Unlike existing gene selection methods, the idea behind EFRL21 and REFRL21 consider global and local information in a unified objective function and feature selection is achieved by adding $\ell_{2,1}$-regularizer term. The existing methods selected differential expression genes based on only one optimization criterion, such as statistical information. Experiments on six publicly available microarray data sets showed that more discriminative features could be found by considering global and local information simultaneously, regardless of the evaluation metrics that were used. Moreover, in the DLBCL and Leukemia data sets, RFS can achieve similar classification results when compared with the proposed methods. However, EFRL21 can achieve the same results with a few genes, this promising result is very significant for biomarker discovering.

In the future, we will further investigate the biological significance of the gene subsets. To the best of our knowledge, pathway analysis [39] and gene regulatory network [40] are all based on differential expression analysis. Consequently, applying our methods to these fields and interpreting the significance from a biological point of view is a natural direction of our future research work. We may also focus primarily on next-generation gene expression data analysis [41]. We are currently conducting research work in this field.

Acknowledgments

This work is supported by the Program for New Century Excellent Talents in University (Grant NCET-10-0365), National Nature Science Foundation of China (Grant 60973082, 11171369, 61202462, 61272395, 61370171, 61300128), the National Nature Science Foundation of Hunan province (Grant 12JJ2041, 13JJ3091), the Planned Science and Technology Project of Hunan Province (Grant 2009FJ3195, 2012FJ2012), A Project Supported by Scientific Research Fund of Hunan Provincial Education Department (Grant14A047, 14C0438, 14B023) and supported by the Fundamental Research Funds for the Central Universities, Hunan University. CX2013A007.

References

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[2] J. E. Elias and S. P. Gygi, "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry," *Nature Methods*, vol. 4, no. 3, pp. 207–214, 2007.

[3] Y. Nannya, M. Sanada, K. Nakazaki, N. Hosoya, L. Wang, A. Hangaishi, M. Kurokawa, S. Chiba, D. K. Bailey, G. C. Kennedy, and S. Ogawa, "A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays," *Cancer Res.*, vol. 65, no. 14, pp. 6071–6079, 2005.

[4] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41–47, 2001.

[5] J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, "Bayesian factor regression models in the large p, small n paradigm," *Bayesian Statist.*, vol. 7, pp. 733–742, 2003.

[6] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Neural Inform. Process. Syst. 16*, 2003, vol. 16, pp. 234–241.

[7] B. Liao, Y. Jiang, W. Liang, W. Zhu, L. Cai, and Z. Cao, "Gene selection using locality sensitive Laplacian score," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 6, pp. 1146–1156, Nov. 2014.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.

[9] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inform. Process. Syst.*, 2005, pp. 507–514.

[10] S. Bandyopadhyay, S. Mallik, and A. Mukhopadhyay, "A survey and comparative study of statistical tests for identifying differential expression from microarray data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 1, pp. 95–115, Jan. 2014.

[11] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 4, pp. 1106–1119, Jul.-Aug. 2012.

[12] S. Niijima and Y. Okuno, "Laplacian linear discriminant analysis approach to unsupervised feature selection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 6, no. 4, pp. 605–614, Oct. 2009.

[13] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 333–342.

[14] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection," in *Proc. 23rd Nat. Conf. Artif. Intell.*, 2008, vol. 2, pp. 671–676.

[15] L.-K. Luo, D.-F. Huang, L.-J. Ye, Q.-F. Zhou, G.-F. Shao, and H. Peng, "Improving the computational efficiency of recursive cluster elimination for gene selection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 1, pp. 122–129, Jan. 2011.

[16] P. A. Mundra and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection," *IEEE Trans. NanoBiosci.*, vol. 9, no. 1, pp. 31–37, Mar. 2010.

[17] C. Ding, D. Zhou, X. He, and H. Zha, "R 1-PCA: Rotational invariant l 1-norm principal component analysis for robust subspace factorization," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 281–288.

[18] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*., vol. 23, no. 11, pp. 1738–1754, Nov. 2012.

[19] X. Niyogi, "Locality preserving projections," in *Proc. Neural Inform. Process. Syst.*, 2004, vol. 16, p. 153.

[20] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*, 2010.

[21] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using l21-norm," in *Proc. 20th ACM Int. Conf. Inform. Knowl. Manage.*, 2011, pp. 673–682.

[22] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[23] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint l2, 1-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, vol. 23, pp. 1813–1821.

[24] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Tech. Univ. Denmark, pp. 7–15, 2008.

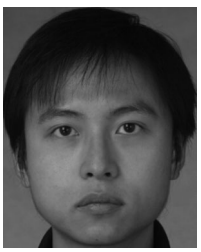
[25] [Online]. Available: http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html, Jul. 2013.

[26] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, C. A. Jeffrey, Z. David, M. O. James, C. Tom, W. Cynthia, A. B. Jaclyn, P. Tomaso, M. Shayan, R. Ryan, C. Andrea, S. Gustavo, N. L. David, P. M. Jill, S. L. Eric, and R. G. Todd, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

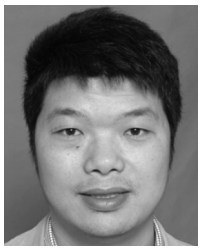
[27] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," in *Proc. Nat. Acad. Sci.*, 1999, vol. 96, no. 12, pp. 6745–6750.

[28] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.
[29] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Med.*, vol. 7, no. 6, pp. 673–679, 2001.
[30] J. D. Spurrier, "On the null distribution of the kruskal–wallis statistic," *Nonparametric Statist.*, vol. 15, no. 6, pp. 685–691, 2003.
[31] F. R. Chung, *Spectral Graph Theory*. American Mathematical Soc., Washington, DC, vol. 92, 1997.
[32] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recog.*, vol. 30, no. 7, pp. 1145–1159, 1997.
[33] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 63–77, Jan. 2006.
[34] Y. Piao, M. Piao, K. Park, and K. H. Ryu, "An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data," *Bioinformatics*, vol. 28, no. 24, pp. 3306–3315, 2012.
[35] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.
[36] S.-L. Wang, Y.-H. Zhu, W. Jia, and D.-S. Huang, "Robust classification method of tumor subtype by using correlation filters," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 2, pp. 580–591, Mar.-Apr. 2012.
[37] X. He, D. Cai, Y. Shao, H. Bao, and J. Han, "Laplacian regularized Gaussian mixture model for data clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 9, pp. 1406–1418, Sep. 2011.
[38] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *Proc.-Int. Joint Conf. Artif. Intell.*, 2011, vol. 22, no. 1, p. 1294.
[39] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, 2009.
[40] D. Husmeier, "Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks," *Bioinformatics*, vol. 19, no. 17, pp. 2271–2282, 2003.
[41] M. Kapushesky, P. Kemmeren, A. C. Culhane, S. Durinck, J. Ihmels, C. Körner, M. Kull, A. Torrente, U. Sarkans, J. Vilo, and A. Brazma, "Expression profiler: Next generationan online platform for analysis of microarray data," *Nucleic Acids Res.*, vol. 32, no. suppl 2, pp. W465–W470, 2004.

Bo Liao received the PhD degree in computational mathematics from the Dalian University of Technology, China, in 2004. He is working at Hunan University as a professor. He worked at the Graduate University of Chinese Academy of Sciences as a post doctorate from 2004 to 2006. His current research interests include bioinformatics, data mining, and machine learning.



Yan Jiang is working towards the PhD degree in the Department of Information Science and Engineering, The Hunan University, China. His research interests include machine learning, pattern recognition, bioinfomatics, and convex optimization.

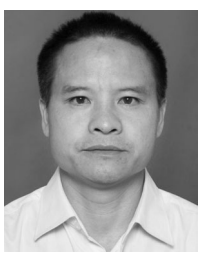


Wei Liang received the BS degree in automation from Central South University, China, in 2001, the MS degree in computer science and technology from the Hunan University of Science and Technology, China, in 2008, and the PhD degree in computer science and technology at the School of Computer and Communication of Hunan University, in China. His current research interests include steganography, steganalysis, real-time embedded systems, intellectual property protection, field programmable gate arrays, and bioinformatics.

Lihong Peng is working towards the PhD degree in the Department of Information Science and Engineering, The Hunan University, China. She is working in Changsha Medical University as a lecturer. Her research interests include machine learning,data mining, and bioinformatics.

Li Peng received the master of engineering degree from the School of Computer Science and Engineering, HuNan University of Science and Technology in 2009. She is currently working towards the PhD degree at Hunan University, China. Her current research focuses on big data, machine learning, and biocomputing.

Damien Hanyurwimfura received the master's degree of engineering in computer science and technology from Hunan University, Changsha, China, in 2010. He is currently working towards the PhD degree at the College of Information Science and Engineering, Hunan University, China. He is also teaching at the College of Science and Technology, University of Rwanda, Rwanda. His current research interests include information security and data mining.



Zejun Li is working towards the PhD degree in the Department of Information Science and Engineering, The Hunan University, China. His research interests include pattern recognition and bioinfomatics.

Min Chen is working towards the PhD degree in the Department of Information Science and Engineering, The Hunan University, China. His research interests include pattern recognition and bioinfomatics.