Journal of Universal Computer Science, vol. 24, no. 6 (2018), 725-741
submitted: 22/5/17, accepted: 15/10/17, appeared: 28/6/18 © *J.UCS*

Unsupervised Feature Selection for Microarray Gene Expression Data Based on Discriminative Structure Learning

Xiucai Ye
(Department of Computer Science, University of Tsukuba
Tsukuba City, Ibaraki, Japan
yexiucai@mma.cs.tsukuba.ac.jp)

Tetsuya Sakurai
(Department of Computer Science, University of Tsukuba
Tsukuba City, Ibaraki, Japan
sakurai@cs.tsukuba.ac.jp)

Abstract: The analysis of microarray gene expression data to obtain useful information is a challenging problem in bioinformatics. Feature selection is an efficient computational technique in processing the analysis of high-dimensional microarray data. Due to the lack of label information in practice, unsupervised feature selection is considered to be more practically important and correspondingly more difficult. In this paper, we propose a novel unsupervised feature selection method, which utilizes local regression and discriminant analysis for structure learning on microarray gene expression data. By imposing row sparsity on the weight matrix through $l_{2,1}$-norm regularization, the proposed method optimizes for selecting the discriminative genes which are more informative and better capture the interesting natural classes of samples. We develop an effective algorithm to solve the $l_{2,1}$-norm-based optimization problem in our method and present the convergence analysis. Finally, we evaluate the proposed method on real microarray gene expression datasets. The experimental results demonstrate that the proposed method not only achieves good performance, but also outperforms other state- of-the-art unsupervised feature selection methods.

Key Words: unsupervised feature selection, structure learning, local regression, discriminant analysis, gene selection, microarray gene expression data.
Category: H.3.2, I.2.6, L.3.2

1 Introduction

In recent decades, DNA microarray techniques have enabled biologists to simultaneously measure the expression level of thousands of genes in specific samples at a given time and under certain conditions [De Rinaldis 2007]. High-throughput expression profiling can be used to compare the level of gene transcription to obtain valuable biological information, and thus assists in diagnosis, prognosis, and treatment of the disease. In a typical microarray dataset, the number of genes is several thousand, far exceeding the limited number of samples, which is one of the main problems for data analysis. Although a large number of genes are measured in experiments, usually many genes are not useful for producing a desired learning or predictive result. The limited number of samples may lead to overfitting due to the noisy genes. Thus, the direct application of data analysis methods to original high-dimensional microarray data is usually inefficient [Somorjai et al. 2003]. In microarray data, usually only a smalle number of genes show strong correlation with the targeted phenotypes [Golub et al. 1999], which are called informative genes. Selecting the informative genes is important for data analysis to identify the set of genes that can further help in finding the biological information embedded in microarray data.

Feature selection has become one of the most important computational techniques to microarray data analysis. The goal of feature selection is to search for the most discriminant feature/gene subset that can distinguish different classes. Feature selection brings the following immediate effects for microarray data analysis: speeding up the algorithms, reducing the risk of overfitting, and improving the accuracy of the predictive results [Dy and Brodley 2004]. Based on the availability of label information, feature selection can be broadly classified into supervised and unsupervised methods [Guyon et al. 2002]. In many bioinformatics applications, with the rapid accumulation of high-dimensional data, the given datasets are usually without any class label information, and it is usually too expensive to perform the labeling through experts [Zhang et al. 2002]. Thus, it is of great importance to develop unsupervised approaches that can perform the feature selection task with only the unlabeled data.

In this paper, we propose a novel method for unsupervised feature selection, which incorporates local regression and discriminant analysis into a learning model to select genes in microarray data analysis. The global structure of microarray data is captured by discriminant analysis, and the local manifold structure is revealed by local regression. $l_{2,1}$-norm sparse regression is also added in the model as a constraint to learn the gene weights correlatively. The resultant formulation of the proposed method optimizes for selecting the most discriminative features which can better capture both the global and local data structure, i.e., selecting the most discriminative genes that are more informative and better capture the interesting natural clusters of samples. We develop an iterative algorithm to effectively solve the optimization problem in the proposed method. We also present the convergence analysis of the algorithm. Experimental results on six real microarray gene expression datasets demonstrate the effectiveness of the proposed method.

2 Related work

Many studies have addressed supervised feature selection by learning with a training set where there are samples with known class labels [Hall 2000, Yassein et al. 2016]. Unsupervised feature selection is more challenging than supervised feature selection, since the definition of relevance of features becomes unclear due to the lack of label information [Dy and Brodley 2000].

Unsupervised feature selection has attracted increasing attention in recent years. Without the label information, a variety of methods have been adopted to perform unsupervised feature selection by extracting features that effectively maintain the important underlying structure of data [Ye et al. 2016]. The global data structure is quite important in data structure learning. The Max Variance (MaxVar) method [Herrero et al. 2003] and principal component analysis [Ding 2003] are two typical methods to preserve the global structure of data.

Instead of learning the global structure, a family of methods perform unsupervised feature selection by preserving the local structure of data. The importance of local structure learning has been well recognized in the recent development of unsupervised feature selection methods. LapScore considers the local preserving property of individual features [He et al. 2006]. MCFS [Cai et al. 2010] selects the features that can best preserve the multi-cluster structure. JELSR [Hou et al. 2011] constructs a graph based on locally linear approximation, and unifies embedding learning with sparse regression to perform feature selection. NDFS [Liu et al. 2012] utilizes spectral clustering to learn the cluster labels, by which a feature subset is selected during the learning of cluster labels.

Discriminant analysis is important to unsupervised feature selection. The objective of discriminant analysis is to select discriminative features such that the within-class distance is as small as possible and the between-class distance is as large as possible [Fukunaga 2013] . Yang et al. [Yang et al. 2012] have proposed a local discriminant analysis method for unsupervised feature selection by defining a local discriminative score to evaluate the within-class scatter and the between-class scatter. However, the discriminant analysis is applied only to local structure learning, which may neglect some informative features from the global perspective.

In this paper, we consider both global and local structure learning for unsupervised feature selection in Microarray data analysis. We apply discriminant analysis for global structure learning. Meanwhile, we utilize local regression for local structure learning. Local regression is effective for capturing the nonlinear geometrical information of data [Sun et al. 2008]. We incorporate local regression, discriminant analysis and $l_{2,1}$-norm regularization into a framework for unsupervised feature learning, with the objective to select the most discriminative genes in microarray data analysis.

3 Notations and preliminaries

In a gene expression microarray study, the output of the gene expression microarray study is recorded as a gene expression data matrix $X = (x_{ij})_{m\times n}$ $(m \gg n)$ that contains the expression of m features/genes across n samples. For the sake of convenience, we use s_1, s_2, ..., $s_n$to denote the n unlabeled samples and g_1, g_2,..., g_m to denote the m genes. Thus, x_{ij} is the expression level of gene g_i in sample s_j.

Feature selection is to select the most informative d $(d < m)$ genes to differentiate the samples originating from different clusters. Consider that s_1, s_2, ..., s_n are sampled from c clusters. We use $L = [l_1, 1_2, ..., l_c] \in \{0, 1\}^{n\times c}$ to denote the label matrix, where $l_i = [l_{1i}, l_{2i}, ..., l_{ni}]^T \in \{0, 1\}^{n\times 1}$ is the label vector containing the labels of n

samples in cluster i. $l_{ji} = 1$ if s_j is in cluster i, and $l_{ji} = 0$ otherwise. Let F denote the scaled cluster indicator matrix and define it as $F = [F_1, F_2, ..., F_n]^T = L(L^T L)^{-1/2}$.

3.1 Local regression

We construct a local label predictor $p_i = [p_{1i}, p_{2i}, ..., p_{ni}]^T$ to estimate the cluster label l_i in cluster i. We choose kernel regression as the local predictor. Motivated by the kernel density estimation [Sun et al. 2008], based on the neighborhood of s_i, the predictor $p_{ti}(t = 1, .., n)$ is defined as

$$p_{ti} = \frac{\sum_{s_j \in N_i} K(s_i, s_j) l_{tj}}{\sum_{s_j \in N_i} K(s_i, s_j)}, \tag{1}$$

where $K(\cdot, \cdot)$ is the kernel function and N_i is the neighborhood of s_i. Define a matrix $M = (m_{ij})_{n \times n}$, where

$$m_{ij} = \begin{cases} \frac{K(s_i, s_j)}{\sum_{s_j \in N_i} K(s_i, s_j)}, & s_j \in N_i, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

From equations (1) and (2), we have $p_i = M l_i$. Let $P = [p_1, p_2, ..., p_c]$ be the predictive matrix. Thus, we can obtain $P = ML$. To reduce the side effect of irrelevant and noisy genes, we formulate the optimization problem by using l_1-norm regularization which has the effect of reducing the large fitting error.

$$\min_L \|L - ML\|_1 = \min_{l_i} \sum_{i=1}^{c} \|l_i - M l_i\|_1. \tag{3}$$

As suggested in [Sun et al. 2008], we use the scaled cluster indicator matrix F to replace L, since equation (3) is difficult to derive in a quadratic form, which can lead to better performance in practice.

$$\min_F \|F - MF\|_1. \tag{4}$$

From [Sun et al. 2008], we know that equation (4) is equivalent to minimizing the following problem,

$$\min_F Tr(F^T G F), \tag{5}$$

where $G = B - (M + M^T)$ and B is an $n \times n$ diagonal matrix with $b_i = \sum_{i=1}^{n} (M + M^T)_{ij}$ on the diagonal. Since $F = L(L^T L)^{-1/2}$, it can be proved that $F^T F = I_c$, where I_c is an identity matrix with c dimensions.

3.2 Discriminant analysis

The linear discriminant analysis is to find a linear transformation matrix W that projects the data matrix X to the low-dimensional space $W^T X$. The total scatter matrix S_t and the between-cluster scatter matrix S_b are defined as [Ye et al. 2016, Fukunaga 2013]

$$S_t = \sum\nolimits_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T = \tilde{X}\tilde{X}^T, \tag{6}$$

$$S_b = \sum\nolimits_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)^T = \tilde{X}FF^T\tilde{X}^T, \tag{7}$$

where μ is the mean of all data, μ_i is the mean of data in cluster i, n_i is the number of data in cluster i, $\tilde{X} = XH_n$ is the data matrix after being centered and $H_n = I - \frac{1}{n}1_n 1_n^T$. The objective of linear discriminant analysis is to minimize the within-cluster distance while maximizing the between-cluster distance in the lower dimensional space as

$$\max_{W} Tr((W^T S_t W)^{-1} W^T S_b W). \tag{8}$$

Let $W = [w_1, ..., w_m]^T \in \mathbb{R}^{m \times q}$, where w_i is the i^{th} row of W. In the lower dimensional space $W^T X$, w_i corresponds to the weight of feature g_i. Thus, the weight of each gene can be calculated according to the transformation matrix W.

4 Unsupervised feature learning for gene selection

In this section, we propose a novel method for gene selection in microarray data analysis by unsupervised feature learning. The proposed method incorporates Local regression and Discriminant analysis for unsupervised Feature Selection. Thus, we refer to it as the LDFS method.

4.1 The objective function

By incorporating local regression, discriminant analysis and $l_{2,1}$-norm regularization into a framework for unsupervised feature learning, the objective function of the proposed LDFS method is formulated as

$$\begin{aligned} \min_{W,F} & -Tr(W^T S_b W) + \alpha\|W\|_{2,1} + \beta Tr(F^T G F), \\ & s.t. F^T F = I_c, F \geq 0, W^T S_t W = I, \end{aligned} \tag{9}$$

where α and β are two balanced parameters. The condition of $F = Y(Y^T Y)^{-1/2}$ is relaxed to $F^T F = I_c$. F is constrained to be nonnegative, which can help to relieve the deviation from the true solution [Liu et al. 2012]. To avoid the trivial solution [Tao et al. 2016], the transformation matrix W is constrained to be uncorrelated with respect to S_t, i.e., $W^T S_t W = I$.

The term $\|W\|_{2,1}$ in equation (9) is introduced to ensure that the transformation matrix W is sparse in rows. Since the i^{th} row w_i corresponds to the weight of gene g_i, the sparsity constraint on rows make W suitable for gene selection. Each gene is ranked according to $\|w_i\|_2$ in descending order and the top d genes are selected.

4.2 Optimization

The optimization problem in equation (9) is not convex when both W and F are optimized simultaneously. Also, the $l_{2,1}$-norm regularization term is non-smooth. We propose an iterative algorithm to divide the problem into two steps: learning W while fixing F, and learning F while fixing W.

According to equations (6) and (7), we replace S_t with $\tilde{X}\tilde{X}^T$ and replace S_b with $\tilde{X}FF^T\tilde{X}^T$ in equation (9). Furthermore, we replace $F^TF = I_c$ with $\frac{\gamma}{2}\|F^TF - I_c\|_F^2$ in the objective function. Thus, equation (9) can be rewritten as

$$\min_{W,F} -Tr(W^T\tilde{X}FF^T\tilde{X}^TW) + \alpha\|W\|_{2,1} + \beta Tr(F^TGF) + \frac{\gamma}{2}\|F^TF - I_c\|_F^2, \\ s.t. F \geq 0, W^T\tilde{X}\tilde{X}^TW = I, \tag{10}$$

where $\gamma > 0$ is a parameter which should be large enough to ensure the orthogonality.

4.2.1 Optimize *W* by fixing *F*

The optimization problem for updating W is equivalent to the following problem.

$$\min_{W} Tr(W^T\tilde{X}FF^T\tilde{X}^TW) + \alpha\|W\|_{2,1}, \\ s.t. W^T\tilde{X}\tilde{X}^TW = I. \tag{11}$$

Let $U \in \mathbb{R}^{m\times m}$ be a diagonal matrix with the i^{th} diagonal element as $U_{ii} = \frac{1}{2\|w_i\|_2}$. Since $\frac{\partial\|W\|_{2,1}}{\partial W} = 2UW$, by constructing an auxiliary function and replacing $\|W\|_{2,1}$ with W^TUW in equation (11), the problem is equivalent to

$$\min_{W} Tr(W^T(-\tilde{X}FF^T\tilde{X}^T + \alpha U)W), \\ s.t. W^T\tilde{X}\tilde{X}^TW = I. \tag{12}$$

The solution of equation (12) can be obtained by solving the following generalized eigenproblem.

$$(-\tilde{X}FF^T\tilde{X}^T + \alpha U)\tilde{w} = \lambda\tilde{X}\tilde{X}^T\tilde{w}. \tag{13}$$

The matrix $W \in \mathbb{R}^{m\times q}$ which contains the eigenvectors corresponding to the q smallest eigenvalues as the column vectors is the solution of (13). Then, we normalize W such that $(W^T\tilde{X}\tilde{X}^TW)_{ii} = 1, i = 1, ..., q$.

4.2.2 Optimize *F* by fixing *W*

The optimization problem for updating F is equivalent to the following problem.

$$\min_F -Tr(W^T\tilde{X}FF^T\tilde{X}^TW) + \beta Tr(F^TGF) + \frac{\gamma}{2}\|F^TF - I_c\|_F^2, \quad s.t. F \geq 0. \tag{14}$$

Since $Tr(W^T\tilde{X}FF^T\tilde{X}^TW) = Tr(F^T\tilde{X}^TWW^T\tilde{X}F)$, let $Q = \beta G - \tilde{X}^TWW^T\tilde{X}$, and (14) can be rewritten as

$$\min_F Tr(F^TQF) + \frac{\gamma}{2}\|F^TF - I_c\|_F^2, \quad s.t. F \geq 0. \tag{15}$$

Following [Liu et al. 2012], we update F by multiplicative rules, as

$$F_{ij} \leftarrow F_{ij}\frac{(\gamma F)_{ij}}{(MF + \gamma FF^TF)_{ij}}. \tag{16}$$

Then, F is normalized to satisfy that $(F^TF)_{ii} = 1, i = 1, ..., n$.

Based on the above analysis, we summarize the procedure of the proposed LDFS method in Algorithm 1. The proposed algorithm will stop when the objective function of equation (9) tends to a constant or the change is smaller than a threshold. The threshold is set very close to zero.

4.3 Discussion

In this section, we first show the convergence behavior of Algorithm 1 and then discuss the time complexity.

4.3.1 Convergence analysis

We prove that the objective function of equation (9) is non-increasing under the updating rules of W and F in Algorithm 1. Before analysis, we show a lemma from [Nie et al. 2010].

Lemma 1 *For any nonzero vectors s and h, the following inequality holds:*

$$\|s\|_2 - \frac{\|s\|_2^2}{2\|h\|_2} \leq \|h\|_2 - \frac{\|h\|_2^2}{2\|h\|_2}. \tag{17}$$

The convergence behavior of LDFS is summarized in the following theorem.

Theorem 1 *Algorithm 1 will monotonically decrease the value of the objection function in equation (9) in each iteration.*

Algorithm 1 The proposed LDFS method

Input: Gene expression data matrix $X \in \mathbb{R}^{m\times n}$; Parameters α, β, γ, k, c, q; Number of features to select d;
Output: d selected features;
1: Construct the k-nearest neighbor graph and calculate M and G;
2: The iteration step $t = 1$; Initialize $F^1 \in \mathbb{R}^{n\times c}$ and set $U^1 \in \mathbb{R}^{m\times m}$ as an identity matrix;
3: Calculate W^1 by solving the generalized eigenproblem $(-\tilde{X}F^1(F^1)^T\tilde{X}^T + \alpha U^1)\tilde{w} = \lambda\tilde{X}\tilde{X}^T\tilde{w}$;
4: **repeat**
5: Calculate $Q^t = \beta G - \tilde{X}^T W^t (W^t)^T \tilde{X}$;
6: $F_{ij}^{t+1} = F_{ij}^t \frac{(\gamma F^t)_{ij}}{(Q^t F^t + \gamma F^t (F^t)^T F^t)_{ij}}$;
7: Update the diagonal matrix U^{t+1} with the i^{th} diagonal element as $U_{ii}^{t+1} = \frac{1}{2\|w_i^t\|_2}$;
8: Calculate W^{t+1} by solving the generalized eigenproblem $(-\tilde{X}F^{t+1}(F^{t+1})^T\tilde{X}^T + \alpha U^{t+1})\tilde{w} = \lambda\tilde{X}\tilde{X}^T\tilde{w}$;
9: t=t+1;
10: **until** Convergence
11: Sort each gene g_i according to $\|w_i\|_2$ in descending order and select the top d ranked ones.

Proof: We rewrite the formulation in equation (9) as

$$\begin{aligned}\Theta(W,F) = -Tr(W^T\tilde{X}FF^T\tilde{X}^TW) + \alpha\|W\|_{2,1} + \beta Tr(F^TGF) \\ +\frac{\gamma}{2}\|F^TF - I_c\|_F^2.\end{aligned} \tag{18}$$

We show that $\Theta(W^{t+1}, F^{t+1}) \le \Theta(W^t, F^t)$.

We first prove $\Theta(W^{t+1}, F^t) \le \Theta(W^t, F^t)$ when F^t is fixed. With F^t fixed, $\Theta(W^t, F^t) = Tr((W^t)^T B^t W^t) + \alpha\|W^t\|_{2,1}$. In the $(t+1)^{th}$ iteration, W^{k+1} is obtained from

$$\min_{W, W^T\tilde{X}\tilde{X}^TW=I} Tr(W^T(-\tilde{X}F^t(F^t)^T\tilde{X}^T + \alpha U^t)W), \tag{19}$$

which indicates that

$$\begin{aligned}Tr((W^{t+1})^T(-\tilde{X}F^t(F^t)^T\tilde{X}^T + \alpha U^{t+1})W^{t+1}) \\ \le Tr((W^t)^T(-\tilde{X}F^t(F^t)^T\tilde{X}^T + \alpha U^t)W^t).\end{aligned} \tag{20}$$

Since $\|W\|_{2,1} = \sum_{i=1}^{m} \|w_i\|_2$, we obtain

$$\begin{aligned} &Tr((W^{t+1})^T(-\tilde{X}F^t(F^t)^T\tilde{X}^T)W^{t+1}) + \alpha\|W^{t+1}\|_{2,1} \\ +\alpha\sum\nolimits_{i=1}^{m}&(\frac{\|w_i^{t+1}\|_2^2}{2\|w_i^t\|_2} - \|w_i^{t+1}\|_2) \leq Tr((W^t)^T(-\tilde{X}F^t(F^t)^T\tilde{X}^T)W^t) \\ &+\alpha\|W^t\|_{2,1} + \alpha\sum\nolimits_{i=1}^{m}(\frac{\|w_i^t\|_2^2}{2\|w_i^t\|_2} - \|w_i^t\|_2). \end{aligned} \tag{21}$$

According to Lemma 1, we know

$$\frac{\|w_i^{t+1}\|_2^2}{2\|w_i^t\|_2} - \|w_i^{t+1}\|_2 \geq \frac{\|w_i^t\|_2^2}{2\|w_i^t\|_2} - \|w_i^t\|_2. \tag{22}$$

Combining equations (21) and (22), we have

$$\begin{aligned} &Tr((W^{t+1})^T(-\tilde{X}F^t(F^t)^T\tilde{X}^T)W^{t+1}) + \alpha\|W^{t+1}\|_{2,1} \\ &\leq Tr((W^t)^T(-\tilde{X}F^t(F^t)^T\tilde{X}^T)W^t) + \alpha\|W^t\|_{2,1} \end{aligned} \tag{23}$$

That is

$$\Theta(W^{t+1}, F^t) \leq \Theta(W^t, F^t). \tag{24}$$

Next, we prove $\Theta(W^t, F^{t+1}) \leq \Theta(W^t, F^t)$ when W^t is fixed by using the method in [Yang et al. 2011]. For the sake of convenience, we denote

$$g(F) = Tr((F^TQF) + \frac{\gamma}{2}\|F^TF - I_c\|_F^2. \tag{25}$$

With W^t fixed, we have $\Theta(W^t, F^t) = g(F^t)$. It is easy to prove $g(F^{t+1}) \leq g(F^t)$. Thus, we have

$$\Theta(W^t, F^{t+1}) \leq \Theta(W^t, F^t). \tag{26}$$

According to equation (24), we have $\Theta(W^{t+1}, F^{t+1}) \leq \Theta(W^t, F^{t+1}) \leq \Theta(W^t, F^t)$. Thus, Algorithm 1 monotonically decreases the objective value in each iteration till convergence. □

4.3.2 Complexity analysis

To optimize the objective function of LDFS, the most time-consuming operation is to solve the generalized eigenproblem $(-\tilde{X}FF^T\tilde{X}^T + \alpha U)\tilde{w} = \lambda\tilde{X}\tilde{X}^T\tilde{w}$, which has a time complexity of $O(m^3)$, where m is the number of features/genes. Empirical results show that the convergence is fast and only several iterations (less than 10 iterations in the experiments) are needed to reach convergence. Thus, the proposed method scales well in practice.

Table 1: Properties of Datasets

Dataset	# of samples	# of Genes	# of Clusters
LUNG	203	3312	5
COLON	62	2000	2
TOX-171	171	5748	4
GLIOMA	50	4434	4
LYMPHOMA	96	4026	9
ALLAML	72	7129	2

5 Experiment

In this section, we conduct experiments to evaluate the performance of the proposed LDFS method on microarray gene expression datasets. We perform two groups of experiments. In the first group, we test the performance of LDFS by using K-means clustering. In the second group, we test the performance of LDFS by using Nearest Neighbors (NN) classifier. We compare the proposed LDFS method with several state-of-the-art unsupervised feature selection methods, including LapScore [He et al. 2006], MCFS [Cai et al. 2010], JELSR [Hou et al. 2011] and NDFS [Liu et al. 2012]. We also compare these feature selection methods with the baseline method which uses all the features for clustering and classification. In the experiments, the number of selected genes is ranged over $\{20, 40, 60, 80, 100, 120, 140, 160, 180, 200\}$. The parameters are tuned over $\{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2, 10^4, 10^6\}$. The number of nearest neighbors is set as $k = 5$. We report the best result of all the methods by using different parameters.

5.1 Dataset description

In the experiments, six public gene expression datasets are collected to illustrate the performance of different feature selection methods. The datasets are LUNG, COLON, TOX-171, GLIOMA, LYMPHOMA and ALLAML, which are downloaded from `http://featureselection.asu.edu/datasets.php`. We summarize the properties of the datasets in Table 1 and briefly introduce them as follows.

- LUNG: 203 samples are described by the expression level of 12600 genes. The samples consist of 5 clusters with 139, 21, 20, 6, and 17 samples. The 3312 genes with standard deviations larger than 50 expression units are retained for the 203 samples.
- COLON: 62 samples are collected from two clusters. 22 samples in one of the clusters are collected from a tumor biopsy, and the other 40 samples in the other cluster are normal.

Table 2: NMI (mean% $\pm$ std) of Different Methods

Dataset	LUNG	COLON	TOX-171	GLIOMA	LYMPHOMA	ALLAML
Baseline	48.1±3.6	4.3±0.6	10.1±1.6	41.1±2.1	60.4±4.2	9.8±4.4
LapScore	52.7±4.9	11.0±1.8	14.2±2.1	48.5±2.7	65.3±2.7	13.6±4.2
MCFS	58.8±5.2	7.8±0.5	24.4±1.0	47.0±3.1	65.0± 2.9	13.2±5.0
JELSR	60.8±4.2	70.4±1.7	22.8±4.8	49.8±3.3	59.7±3.6	10.7±4.5
NDFS	60.8±4.1	6.1±2.2	24.9±1.0	50.6±2.9	60.8± 3.8	12.2±5.3
LDFS	**63.2±4.4**	**11.9±2.0**	**25.9±1.7**	**51.2±3.0**	**65.8±3.2**	**14.1±4.0**

- TOX-171: 5748 genes are taken from 171 samples. The samples consist of 4 clusters with 45, 45, 39, and 42 samples.
- GLIOMA: 50 samples are collected from 4 clusters with 14, 7, 14, and 15 samples. The samples contain the expression level of 4434 genes.
- LYMPHOMA: 96 samples contain the expression level of 4026 genes. The samples consist of 9 clusters with 46, 11,10, 9, 6, 6, 4, 2, and 2 samples.
- ALLAML: 7129 genes are taken from 72 samples, which belong to patients suffering from acute myeloid leukemia (AML: 25 samples) and acute lymphoblastic leukemia (ALL: 47 samples).

5.2 Results by clustering

In the first group experiment, K-means clustering is applied to evaluate the performance of LDFS. We apply two widely used evaluation metrics, i.e., Normalized Mutual Information (NMI) and Accuracy (ACC), to evaluate the clustering results. $C = \{C_i\}_{i=1}^{c}$ denotes the ground truth clustering configuration of a dataset, where c is the ground truth cluster number. $C' = \{C'_i\}_{i=1}^{c'}$ denotes the clustering configuration obtained by a clustering algorithm, where c' is the obtained cluster number. n is the cardinality of the whole dataset. n_i is the cardinality of C_i. n'_i is the cardinality of C'_i. And, n_{ij} is the cardinality of the intersection of C_i and C'_i. The NMI criteria is defined as

$$NMI(C, C') = \frac{\sum_{i=1}^{c}\sum_{j=1}^{c'} n_{ij}\log(n \cdot n_{ij}/(n_i \cdot n_j))}{\sqrt{(\sum_{i=1}^{c} n_i \log(n_i/n))(\sum_{j=1}^{c'} n_j \log(n_j/n))}}. \tag{27}$$

A larger value of NMI indicates better performance. Let l_i denote the ground truth label of s_i and l'_i denote the index of clustering result of s_i. ACC is defined as [Ye et al. 2016]

$$ACC(C, C') = \frac{\sum_{i=1}^{n} \delta(l_i, map(l'_i))}{n}, \tag{28}$$

Table 3: ACC (mean% $\pm$ std) of Different Methods

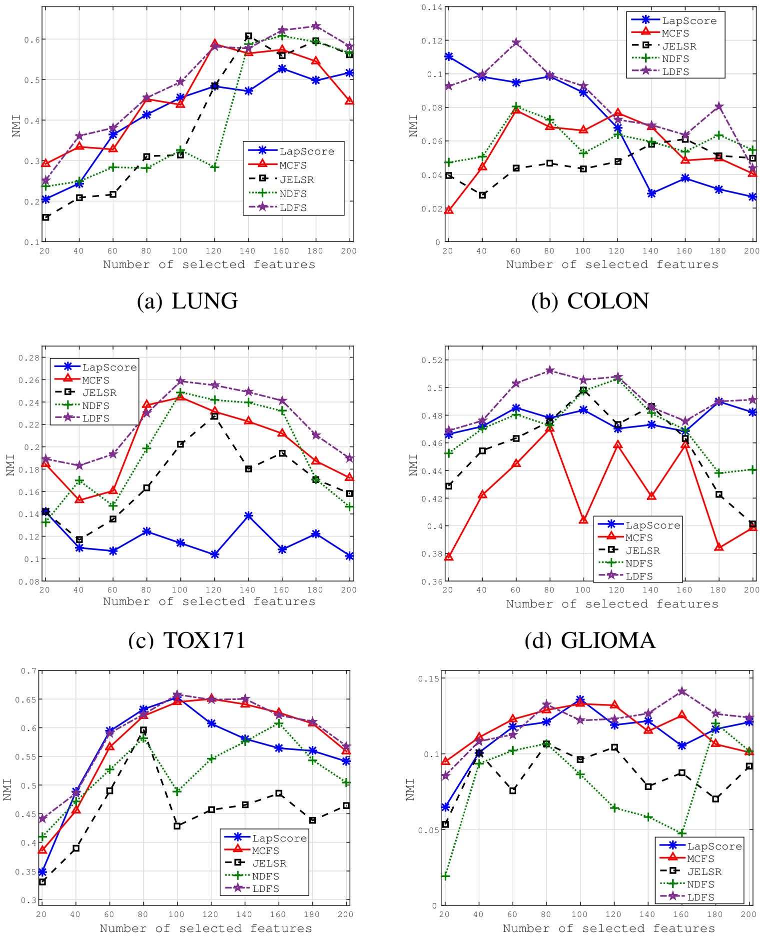
Dataset	LUNG	COLON	TOX-171	GLIOMA	LYMPHOMA	ALLAML
Baseline	67.0±1.6	53.2±2.3	40.9±2.6	52.0±3.1	56.5±5.3	68.2±6.5
LapScore	60.4±1.9	60.5±2.8	43.9±2.1	60.1±3.2	62.5±3.6	73.7±5.1
MCFS	81.3 ±3.2	58.2±2.5	48.1±3.0	60.4±2.7	60.7± 3.8	73.4±5.8
JELSR	78.7±4.2	58.3±3.2	47.2±3.8	59.5±3.4	58.3±4.2	70.5±6.1
NDFS	77.5±4.1	59.4±3.0	47.6±3.1	58.4±3.3	58.6± 4.5	72.6±5.7
LDFS	**85.2±2.4**	**61.3±2.5**	**49.6±3.0**	**61.5±2.9**	**63.2±4.0**	**74.1±5.5**

where $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ otherwise, $map(l'_i)$ is the best mapping function that permutes clustering labels to match the ground truth labels using the Kuhn-Munkres algorithm. A larger value of ACC indicates a better clustering result.

Each feature selection method is first performed to select genes on the gene expression data sets. After selecting the genes, K-means clustering is performed by using only the selected genes. We repeat each experiment 20 times with random initializations and report the mean performance with standard deviation.

First, we compare the performance of the feature selection methods on the six gene expression datasets. The experimental results in terms of NMI and ACC evaluation metrics are shown in Tables 2 and 3, respectively. We can see from the two tables that most of the unsupervised feature selection methods perform better than the baseline method. Gene selection can improve the accuracy of clustering results. The proposed LDFS method performs better than the other methods on the six datasets. This is because LDFS utilizes local regression and discriminative analysis simultaneously to learn the weight of each gene.

Then, we evaluate the performance of the clustering results on the six datasets by varying the number of selected genes. The performance of the clustering results in terms of NMI and ACC evaluation metrics are shown in Figures 1 and 2, respectively. We can see from the figures that the proposed LDFS method performs better than other methods in most cases when selecting different number of genes. Note that for different datasets, the numbers of selected genes to obtain the best results are different. For example, in Figure 1, on the Lung dataset, the optimized gene number is about 180, while on the Colon dataset, the optimized gene number is about 60. This is because in different microarray datasets, the correlations of genes are different. In Figure 2, the trend of the performance when using the ACC evaluation metric is very similar to that when using the NMI evaluation metric. The proposed LDFS method obtains better performance than other methods when both NMI and ACC evaluation metrics are applied.



(a) LUNG (b) COLON

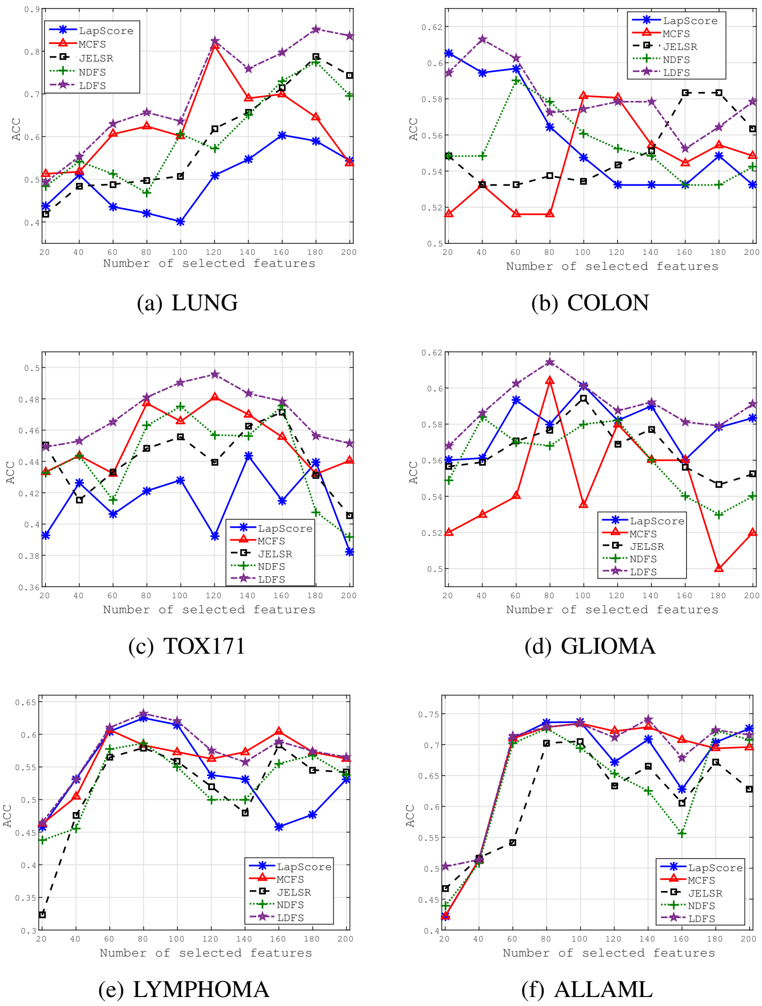
(c) TOX171 (d) GLIOMA

(e) LYMPHOMA (f) ALLAML

Figure 1: NMI by varying the number of selected features/genes.

5.3 Results by classification

In the second group, Nearest Neighbors (NN) classifier is applied to test the performance of LDFS. We utilize 5-fold cross-validation by which the original samples are randomly partitioned into 5 equal-sized subsets. One subset is retained as the validation data for testing the model, and the remaining 4 subsets are used as training data. The cross-validation process is then repeated 5 times (the folds), with each of the 5 subsets of samples being used exactly once as the validation data. We perform gene selection using the training data, and evaluate the performance of the selected features on the test data. The experiments are repeated 20 times on the best parameter combination. We report the mean classification error with standard deviation.



(a) LUNG (b) COLON

(c) TOX171 (d) GLIOMA

(e) LYMPHOMA (f) ALLAML

Figure 2: ACC by varying the number of selected features/genes

Table 4 shows the classification results of different methods on the six datasets. We can see that the proposed LDFS method has a lower classification error than the other methods. Most of the unsupervised feature selection methods perform better than the baseline method, except on the LUNG and GLIOMA datasets. Specially, on the LUNG dataset, only the proposed unsupervised feature selection method performs better than the baseline method. This is because, by the proposed method the selected genes have less redundancy and have a higher accuracy in predictive results.

The classification performances when varying the number of selected genes are

Table 4: Classification Error (mean% $\pm$ std) of Different Methods

Dataset	LUNG	COLON	TOX-171	GLIOMA	LYMPHOMA	ALLAML
Baseline	5.9±1.6	22.4±3.3	39.2±4.6	22.0±3.0	16.7±3.0	18.7±4.6
LapScore	8.9±1.7	16.4±3.0	37.9±4.1	22.2±3.2	15.6±2.3	13.9±3.5
MCFS	6.8±1.2	16.1±2.8	29.5±3.8	19.2±2.8	15.5± 2.6	13.6±4.0
JELSR	7.3±2.0	17.2±3.2	31.8±4.2	20.1±3.1	16.6±2.9	17.3±3.7
NDFS	8.4±1.6	17.6±3.1	34.3±3.7	22.0±2.7	16.1± 3.2	15.3±4.1
LDFS	**5.4±2.0**	**15.3±2.9**	**28.8±4.0**	**18.6±3.0**	**15.0±3.1**	**13.2±3.8**

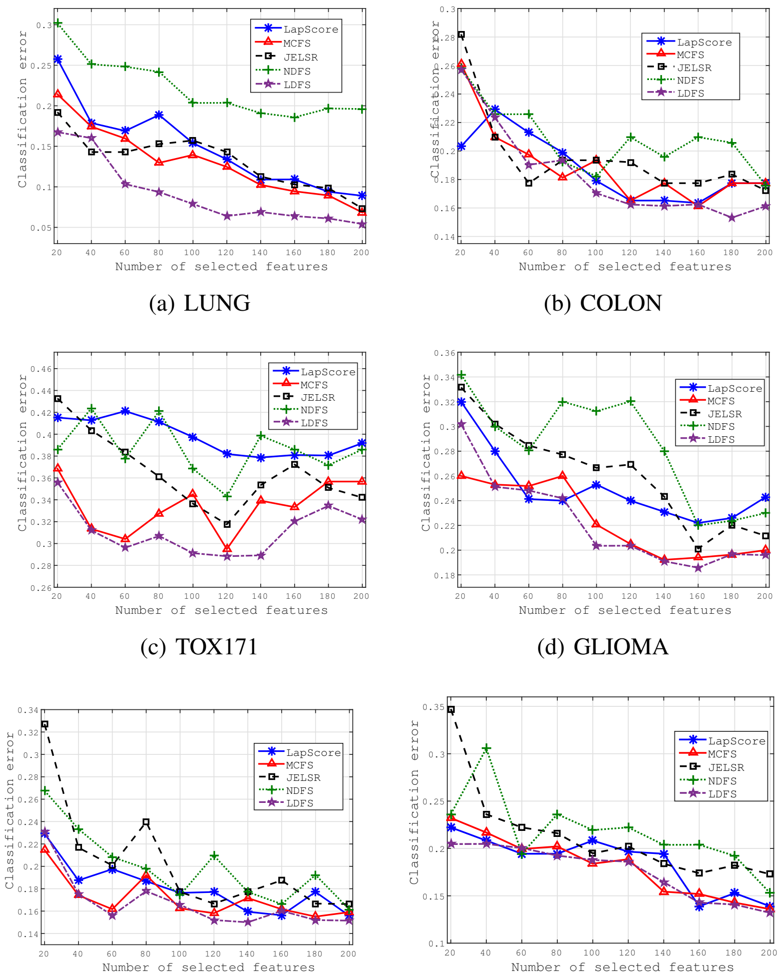
shown in Figure 3. The proposed LDFS method has a lower classification error than other methods on most of the selected features. We note that LDFS also has a better stability than other methods on most of the datasets. Moreover, we also notice that LDFS obtains better results than other methods with fewer features on most of the datasets. For example, on the LUNG, TOX-171, and ALLAML datasets, with the fewest selected features (10 features), LDFS obtains the best performance of all the methods.

6 Conclusions and Future Work

In this paper, we apply unsupervised feature selection in microarray data analysis, by incorporating local regression, discriminant analysis, and $l_{2,1}$-norm regularization into a framework for data structure learning. The proposed method optimizes for selecting the most discriminative genes that have less redundancy and a higher accuracy in predictive results. We derive an effective algorithm to solve the optimization problem of the proposed method and present the convergence analysis. Experiments on six real microarray gene expression datasets demonstrate that the proposed method not only achieves good performance, but also outperforms other state-of-the-art unsupervised feature selection methods. In the experiments, we use the gene expression datasets that are wildly used for feature selection methods to demonstrate their effectiveness. The digital gene expression datasets obtained by next-generation sequencing techniques are interesting and important. We will learn more about the digital gene expression datasets and conduct experiments on the digital gene expression datasets in the further work.

Acknowledgements

This work was supported in part by JST/CREST and MEXT KAKENHI (Grant No.25 286097).



(a) LUNG (b) COLON

(c) TOX171 (d) GLIOMA

(e) LYMPHOMA (f) ALLAML

Figure 3: Classification Error by varying the number of selected features/genes.

References

[Cai et al. 2010] Cai, D., Zhang, C., and He, X.: "Unsupervised feature selection for multi-cluster data"; Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 333-342.

[De Rinaldis 2007] De Rinaldis, E. : "DNA microarrays: current applications"; Horizon Scientific Press, Norfolk.

[Ding 2003] Ding, C.: "Unsupervised Feature Selection via Two-Way Ordering in Gene Expression Analysis"; Bioinformatics. 19, 10(2003), 1259-1266.

[Dy and Brodley 2000] Dy, J. G., and Brodley, C. E.: "Visualization and interactive feature selection for unsupervised data"; Proc. ACM SIGKDD international conference on Knowledge discovery and data mining, 360?364.

[Dy and Brodley 2004] Dy, J. G., and Brodley, C. E.: "Feature selection for unsupervised learning."; Journal of Machine Learning Research 5(2004): 845-889.

[Fukunaga 2013] Fukunaga, K.: 'Introduction to statistical pattern recognition"; Academic press.

[Golub et al. 1999] Golub, T. R., Slonim, D. K. etal.: "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring"; Science. 286, 5439 (Oct, 1999), 531-537.

[Guyon et al. 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V.: "Gene selection for cancer classification using support vector machines"; Machine Learning. 46, 1 (2002), 389-422.

[Hall 2000] Hall, M.: "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning"; Proc. International Conference on Machine Learning, 359- 366.

[He et al. 2006] He, X., Cai, D., and Niyogi, P.: "Laplacian score for feature selection"; Proc. Advances in Neural Information Processing Systems. 507-514.

[Herrero et al. 2003] Herrero, J., Dlaz-Uriarte, R., and Dopazo, J.: "Gene Expression Data Preprocessing"; Bioinformatics. 19, 5 (2003), 655-656.

[Hou et al. 2011] Hou, C., Nie, F., Li, X., Yi, D., and Wu, Y.: "Feature selection via joint embedding learning and sparse regression"; Proc. International Joint Conference on Artificial Intelligence. 1324-1329.

[Liu et al. 2012] Liu, L., Zhou, X., Li, Z., Yang, Y., and Lu, H.: "Unsupervised feature selection using nonnegative spectral analysis"; Proc. AAAI Conference on Artificial Intelligence. AAAI, 1026-1032.

[Nie et al. 2010] Nie, F., Huang, H., Cai, X., and Ding, C.: 'Efficient and robust feature selection via joint l2,1-norms minimization"; Proc. Advances in Neural Information Processing Systems. 1813-1821.

[Somorjai et al. 2003] Somorjai, R.L., Dolenko, B., and Baumgartner, R.: "Class Prediction and Discovery Using Gene Microarray and Proteomics Mass Spectroscopy Data: Curses, Caveats, Cautions"; Bioinformatics. 19, 12 (2003), 1484-1491.

[Sun et al. 2008] Sun, J., Shen, Z., Li, H., and Shen, Y: "Clustering via local regression'; Proc. European conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 456-471.

[Tao et al. 2016] Tao, H., Hou, C., Nie, F., Jiao, Y., and Yi, D.: "Effective discriminative feature selection with nontrivial solution"; IEEE Transactions on Neural Networks and Learning Systems. 27, 4 (Apr.2016), 796-808.

[Yang et al. 2011] Yang, Y., Shen, H. T., Nie, F., Ji, R., and Zhou, X.: 'Non-negative spectral clustering with discriminative regularization"; Proc. AAAI Conference on Artificial Intelligence. 1813-1821.

[Yang et al. 2012] Yang, Y., Shen, H., Ma, Z., Huang, Z., and Zhou, X.: "L2,1-norm regularized discriminative feature selection for unsupervised learning'; Proc. International Joint Conference on Artificial Intelligence. 1589-1594.

[Yassein et al. 2016] Yassein,M. B., Khamayseh, Y., and AbuJazoh, M.: "Feature Selection for Black Hole Attacks "; Journal of Universal Computer Science. 22, 4 (2016), 521-536.

[Ye et al. 2016] Ye, X., Ji, K., and Sakurai, T.: "Unsupervised Feature Selection with Correlation and Individuality Analysis"; International Journal of Machine Learning and Computing. 6,1(February. 2016), 36-41.

[Ye et al. 2016] Ye, X., Ji, K., and Sakurai, T.: "Global Discriminant Analysis for Unsupervised Feature Selection with Local Structure Preservation'; Proc. International Florida Artificial Intelligence Research Society Conference. AAAI, Florida, 454-459.

[Zhang et al. 2002] Zhang, S., Wong, HS., Shen, Y., and Xie, D.: "A New Unsupervised Feature Ranking Method for Gene Expression Data Based on Consensus Affinity"; IEEE/ACM Transactions on Computational Biology and Bioinformatics. 9, 4 (July, 2012) , 1257-1263.