IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON FEATURE REPRESENTATION AND LEARNING METHODS WITH APPLICATIONS IN LARGE-SCALE BIOLOGICAL SEQUENCE ANALYSIS

Received July 31, 2020, accepted August 16, 2020, date of publication August 21, 2020, date of current version September 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3018480

Adaptive Unsupervised Feature Learning for Gene Signature Identification in Non-Small-Cell Lung Cancer

XIUCAI YE, (Member, IEEE), WEIHANG ZHANG, AND TETSUYA SAKURAI, (Member, IEEE)
Department of Computer Science, University of Tsukuba, Tsukuba 3058577, Japan

Corresponding author: Xiucai Ye (yexiucai@cs.tsukuba.ac.jp)

This work was supported in part by the New Energy and Industrial Technology Development Organization 265 (NEDO) and in part by the Japan Society for the Promotion of Science (JSPS), Grants-in-Aid for Scientific Research under Grant 18H03250.

ABSTRACT Non-small-cell lung cancer (NSCLC) is the most common type of lung cancer, which accounts for a proportion of nearly 85%. The increasing availability of genome-wide gene expression data has facilitated the identification of gene signatures that are significant to the precise classification of NSCLC subtypes and personalized treatment decisions. Unsupervised feature selection is an effective computational technique for searching the most discriminative feature subset to distinguish different classes and find the potential information embedded in biological data. In this study, we proposed a novel unsupervised feature selection method to identify the gene signatures for NSCLC subtype classification based on gene expression data. The proposed method incorporated linear discriminant analysis, adaptive structure preservation, and $l_{2,1}$-norm sparse regression into a joint learning framework for unsupervised feature selection to select the informative genes. An effective algorithm was developed to solve the optimization problem in the proposed method. Furthermore, we performed module-based gene filtering before feature selection to reduce the computational cost. We evaluated the proposed method on a gene expression dataset of NSCLC from The Cancer Genome Atlas (TCGA). The experimental results show that the proposed method identified a small number of gene signatures for accurate NSCLC subtype classification. Enrichment analysis of the identified gene signatures was also performed by summarizing the key biological processes.

INDEX TERMS Unsupervised feature selection, non-small-cell lung cancer, subtype classification.

I. INTRODUCTION

Lung cancer which is a highly lethal malignant disease has become the leading cause of cancer-related death worldwide [1]. Small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC) are the two main types of lung cancer, where NSCLC accounts for a proportion of nearly 85% and has a better prognosis than SCLC [2]. Despite recent therapeutic advances, due to the lack of predictive biomarkers, patients with NSCLC still suffer from bleak outcomes [3]. Many studies have indicated that precise treatment can improve the overall survival of individuals with NSCLC if the subtypes can be identified correctly. It is important to develop novel prognostic biomarkers for subtype classification and treatment optimization in NSCLC.

Recent advances in genome-wide sequencing techniques have enabled the generation of a large amount of gene expression profiles, which facilitate the identification of gene signatures that are significant to the precise classification of NSCLC subtypes and personalized treatment decisions [4]. Many existing studies have addressed to extract gene expression signatures for NSCLC subtype classification including mRNA, miRNA, and lncRNA signatures [5]–[7]. However, directly using the gene profiles as the signatures for NSCLC lung cancer subtype classification usually plagued with redundant information [8].

To reduce the redundant information and remove the useless genes, many feature ranking methods have been applied to select the effective gene signatures [9]–[12]. T-score [13] and Relief-F [14] are two typical feature ranking methods, which consider the genes as individual features and select a set of signatures from the top-ranking for cancer subtype classification. Recursive Feature Elimination (RFE) is a popular feature selection algorithm, which is commonly used with Support Vector Machines (SVM), i.e., SVM-RFE method, to repeatedly construct a model and remove

The associate editor coordinating the review of this manuscript and approving it for publication was Leyi Wei.

features with low weights [15]. Some variants of SVM-RFE, e.g., by integrating gene expression level and correlation [16] or pathway knowledge [17], have been proposed to alter the ranking criterion [18]. Other method using both expression and network information also has been developed to identify genes that are better indicators for survival [19]. All the above feature ranking methods are supervised, which need class labels or related gene information to select the effective gene signatures for cancer subtype classification.

Unsupervised learning has attracted much attention in recent years, which looks for previously undetected patterns in data without pre-existing labels. Clustering is one of the main methods of unsupervised learning to discover useful information hidden in data [20]. Similar to the clustering methods, unsupervised feature selection has also been utilized to select the informative features/genes which better capture the interesting natural classes of samples [21], [22].

Two of the widely used unsupervised feature selection methods are the filter and embedded methods. The filter methods are simple and fast but ignore the possible feature correlations. Typical filter methods include the max variance (MaxVar) method and the Laplacian score (LapScore) method [23]. In contrast to the filter methods, the embedded methods consider the correlation of features with a learning model simultaneously. A family of methods has been developed to maintain the underlying data structure in the embedded learning processes [24]. These important structures include the global structure [25], [26], the local structure [27], [28], and the discriminative information [29], [30]. However, in most existing unsupervised feature selection methods, the calculation of preserved structures in the embedded space involves all the irrelevant and relevant features, thus the irrelevant features will have adverse effects on the structure characterization for selecting the precise features.

In this study, to select the effective and precise gene signatures for the NSCLC subtype classification, we proposed a novel unsupervised feature selection method which maintains the important data structure by using only the selected features. The proposed method incorporated linear discriminant analysis, adaptive structure preservation, and $l_{2,1}$-norm sparse regression into a joint learning framework for unsupervised feature learning to select the informative genes. In the proposed method, the global structure was captured by the discriminant analysis, and the local structure was revealed by a probabilistic neighborhood graph using only the relevant features. By utilizing $l_{2,1}$-norm regularization to impose row sparsity on the weight matrix, the proposed method optimized for selecting the discriminative genes that were informative for the NSCLC subtype classification. We developed an effective algorithm to solve the optimization problem in the proposed method. Furthermore, module-based gene filtering was performed before unsupervised feature selection to reduce the computational cost. We evaluated the proposed method on a gene expression dataset of NSCLC from The Cancer Genome Atlas (TCGA). The experimental results demonstrate that the proposed method identified a small number of gene signatures for accurate NSCLC subtype classification. We also performed an enrichment analysis of the selected gene signatures by summarizing the key biological processes.

II. ADAPTIVE UNSUPERVISED FEATURE SELECTION

A. NOTATIONS

The gene expression dataset is recorded as a data matrix $X = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^{n\times m}$, where $x_i \in \mathbb{R}^m$ denotes the ith sample and n is the number of samples. By using $g_1, g_2, \ldots, g_m$ to denote the m genes, the data matrix can also be denoted as $X = [g_1, g_2, \ldots, g_m]$. Unsupervised feature selection is performed with the objective to select the d ($d < m$) most informative genes that can distinguish the samples originating from different classes.

Assume that the n samples are from c classes. Denote $L = [l_1, l_2, \ldots, l_c] \in \{0, 1\}^{n\times c}$ as the label matrix, where $l_i = [l_{1i}, l_{2i}, \ldots, l_{ni}]^T \in \{0, 1\}^{n\times 1}$ is a label vector related to class i, i.e., $l_{ji} = 1$ if x_j is in class i and $l_{ji} = 0$ otherwise. The scaled class indicator matrix is defined and calculated as $F = [F_1, F_2, \ldots, F_n]^T = L(L^TL)^{-1/2}$.

For a matrix $W = (w_{ij}) \in \mathbb{R}^{v\times u}$, the $l_{2,1}$-norm of W is defined as

$$\|W\|_{2,1} = \sum_{i=1}^{v} \sqrt{\sum_{j=1}^{u} w_{ij}^2}. \tag{1}$$

B. LINEAR DISCRIMINANT ANALYSIS

Linear discriminant analysis is to project the data matrix X to a low-dimensional space by a linear transformation matrix W. Thus, the data matrix is transformed to W^TX in the low-dimensional space. Let $W = [w_1, \ldots, w_m]^T \in \mathbb{R}^{m\times q}$, where w_i is the i^{th} row of W. The total scatter matrix S_t and the between-cluster scatter matrix S_b are defined as [31]

$$S_t = \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T = \tilde{X}\tilde{X}^T, \tag{2}$$

$$S_b = \sum_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)^T = \tilde{X}FF^T\tilde{X}^T, \tag{3}$$

where μ is the mean of the n samples, μ_i is the mean of the samples in class i, n_i is the number of samples in class i, $\tilde{X} = XH_n$ is the centered data matrix of X by $H_n = I - \frac{1}{n}1_n1_n^T$.

By minimizing the within-cluster distance and maximizing the between-cluster distance in the lower dimensional space, the objective function of linear discriminant is calculated as

$$\max_W Tr((W^TS_tW)^{-1}W^TS_bW). \tag{4}$$

C. ADAPTIVE STRUCTURE PRESERVATION

Most existing methods preserve the local structure by constructing a k-nearest neighbor graph. However, the k-nearest neighbor graph is constructed based on all the irrelevant and relevant features, which make the captured local structure to be inevitably affected by the irrelevant features. We considered to preserve the local structure by constructing a probabilistic neighborhood graph based on only the relevant features.

Similar to the methods in [32], [33], we used the Euclidean distance to calculate the probabilistic neighborhood. Assume that x_i is connected to x_j with probability p_{ij}, $0 \le p_{ij} \le 1$. The probabilities of all samples being connected to x_i satisfy $\sum_{j=1}^{n} p_{ij} = 1$. Let $P = (p_{ij})_{n\times n}$. p_{ij} can be calculated by solving the following optimization problem.

$$\min_{0\le p_{ij}\le 1, \sum_{j=1}^{n} p_{ij}=1} \sum_{j=1}^{n} (\|x_i - x_j\|^2 p_{ij} + \lambda p_{ij}^2), \tag{5}$$

where λ is the regularization parameter. The regularization term is used to add a prior uniform distribution and avoid the trivial solution. Note that a small distance $\|x_i - x_j\|^2$ will lead to a high probability. Thus, p_{ij} should be large if $\|\phi(x_i) - \phi(x_j)\|^2$ is small. With such a nice property, the probability neighborhood can be used for local structure preservation. That is, in the low- dimensional space $W^T X$, the probabilistic neighborhood can be preserved. Thus, we have

$$\min_{0\le p_{ij}\le 1, \sum_{j=1}^{n} p_{ij}=1} \sum_{j=1}^{n} (\|W^T x_i - W^T x_j\|^2 p_{ij} + \lambda p_{ij}^2). \tag{6}$$

D. PROPOSED METHOD

We proposed a novel unsupervised feature selection method to select the informative genes for accurate NSCLC subtype classification. The proposed method utilized adaptive structure preservation for unsupervised feature selection. Thus, we referred to it as the AUFS method.

By incorporating linear discriminant analysis, adaptive structure preservation, and $l_{2,1}$-norm sparse regression into a learning framework, the objective function of the proposed AUFS method can be formulated as

$$\begin{aligned} \min_{W,F,P} \; & -Tr(W^T S_b W) + \alpha \|W\|_{2,1} \\ & + \beta \sum_{j=1}^{n} (\|W^T x_i - W^T x_j\|^2 p_{ij} + \lambda p_{ij}^2) \\ s.t.\ & F \ge 0, F^T F = I_c, W^T S_t W = I, 0 \le p_{ij} \le 1, \\ & \sum_{j=1}^{n} p_{ij} = 1, \end{aligned} \tag{7}$$

where α and β are two balanced parameters. F is constrained to be nonnegative [25], and the condition of $F = L(L^T L)^{-1/2}$ is relaxed to $F^T F = I_c$. $W^T S_t W = I$ is set to avoid the trivial solution by constraining W to be uncorrelated with respect to S_t [34].

The term $\|W\|_{2,1}$ in (7) is set to ensure that W is sparse in rows. The i^{th} row w_i corresponds to the weight of gene g_i. Thus, the sparsity constraint on rows makes W suitable for gene selection [31]. Each gene is ranked according to $\|w_i\|_2$ in descending order and the top genes will be selected.

E. OPTIMIZATION ALGORITHM

It is difficult to derive the closed solution of the optimization problem in (7) directly, since it contains three different variables with different regularizations and constraints. We developed an iterative algorithm to convert the problem into three sub-problems by updating one variable while fixing the other two variables.

We replaced S_t with $\tilde{X}\tilde{X}^T$ and replaced S_b with $\tilde{X}FF^T\tilde{X}^T$ in (7) according to (2) and (3), respectively. We also replaced $F^T F = I_c$ with $\frac{\gamma}{2}\|F^T F - I_c\|_F^2$. Therefore, equation (7) can be rewritten as

$$\begin{aligned} \min_{W,F,P} \; & -Tr(W^T \tilde{X} F F^T \tilde{X}^T W) + \alpha \|W\|_{2,1} \\ & + \beta \sum_{j=1}^{n} (\|W^T x_i - W^T x_j\|^2 p_{ij} + \lambda p_{ij}^2) \\ & + \frac{\gamma}{2}\|F^T F - I_c\|_F^2, \\ s.t.\ & F \ge 0, W^T \tilde{X}\tilde{X}^T W = I, 0 \le p_{ij} \le 1, \sum_{j=1}^{n} p_{ij} = 1, \end{aligned} \tag{8}$$

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

1) UPDATE W BY FIXING F AND P

Let $L_s = H_n F F^T H_n^T$ and denote $L_p = D_p - (P + P^T)/2$, where D_P is a diagonal matrix with the ith diagonal element being $\sum_j (p_{ij} + p_{ji})/2$. Let $L = \beta L_p - L_s$. When F and P are fixed, the optimization problem in (8) for updating W is equivalent to the following problem.

$$\begin{aligned} \min_{W} \; & Tr(W^T X L X^T W) + \alpha \|W\|_{2,1}, \\ s.t.\ & W^T \tilde{X}\tilde{X}^T W = I. \end{aligned} \tag{9}$$

Denote $U \in \mathbb{R}^{m\times m}$ as a diagonal matrix with the i^{th} diagonal element being $U_{ii} = \frac{1}{2\|w_i\|_2}$. Since $\frac{\partial \|W\|_{2,1}}{\partial W} = 2UW$, we constructed an auxiliary function and replace $\|W\|_{2,1}$ with $W^T U W$ in (9), the problem is equivalent to

$$\begin{aligned} \min_{W} \; & Tr(W^T (X L X^T + \alpha U) W), \\ s.t.\ & W^T \tilde{X}\tilde{X}^T W = I. \end{aligned} \tag{10}$$

The optimization problem of (10) can be solved by the following generalized eigenproblem.

$$(XLX^T + \alpha U)\tilde{w} = \lambda \tilde{X}\tilde{X}^T \tilde{w}. \tag{11}$$

The solution of (10) is the matrix $W \in \mathbb{R}^{m\times q}$ in which the column vectors are the q smallest eigenvectors corresponding to the q smallest eigenvalues in (11). We further normalized W as $(W^T \tilde{X}\tilde{X}^T W)_{ii} = 1, i = 1, \ldots, q$.

2) UPDATE F BY FIXING W AND P

When W and P are fixed, updating F is equivalent to the following problem.

$$\begin{aligned} \min_{F} \; & -Tr(W^T \tilde{X} F F^T \tilde{X}^T W) + \frac{\gamma}{2}\|F^T F - I_c\|_F^2, \\ s.t.\ & F \ge 0. \end{aligned} \tag{12}$$

Since $Tr(W^T \tilde{X} F F^T \tilde{X}^T W) = Tr(F^T \tilde{X}^T W W^T \tilde{X} F)$, (12) can be rewritten as

$$\begin{aligned} \min_{F} \; & -Tr(F^T \tilde{X}^T W W^T \tilde{X} F) + \frac{\gamma}{2}\|F^T F - I_c\|_F^2, \\ s.t.\ & F \ge 0. \end{aligned} \tag{13}$$

Following [24], by denoting $M = -\tilde{X}^T WW^T \tilde{X}$, F can be updated by multiplicative rules, as

$$F_{ij} \leftarrow F_{ij} \frac{(\gamma F)_{ij}}{(MF + \gamma FF^T F)_{ij}}. \tag{14}$$

Then, we normalized F to satisfy $(F^T F)_{ii} = 1, i = 1, \ldots, n$.

3) UPDATE P BY FIXING W AND F

When W and F are fixed, updating P in (8) is equivalent to the optimization problem in (6). Let $g_{ij} = \|W^T x_i - W^T x_j\|^2$. Let $g_i \in \mathbb{R}^{n\times 1}$ denote a vector with the j^{th} element being g_{ij}. Let $p_i \in \mathbb{R}^{n\times 1}$ denote a vector with the j^{th} element being p_{ij}. Let $1_n \in \mathbb{R}^{n\times 1}$ denote a vector with all of its elements being 1. The vector form of (6) can be written as

$$\min_{0\le p_{ij}\le 1, p_i^T 1_n = 1} \|p_i + \frac{g_i}{2\lambda}\|^2. \tag{15}$$

The Lagrangian function of (15) is

$$\Gamma = \frac{1}{2}\|p_i + \frac{g_i}{2\lambda}\|^2 - \mu(p_i^T 1_n - 1) - \sigma_i^T p_i, \tag{16}$$

where μ and σ_i are the Lagrangian multipliers. Based on the KKT condition [35], we can obtain the optimal solution p_{ij} as

$$p_{ij} = (-\frac{g_{ij}}{2\lambda} + \mu)_+. \tag{17}$$

Following [32], [33] and assume that $g_{i1}, g_{i2}, \ldots, g_{in}$ are ordered from large to small, to satisfy that each sample has only k nearest neighbors, the regularization parameter λ can be set based on the number of nearest neighbors k, as

$$\lambda = \frac{k}{2} g_{i,k+1} - \frac{1}{2}\sum_{j=1}^{k} g_{ij}. \tag{18}$$

Compared to the parameter λ, k is much easier and more intuitive to tune. Thus, λ can be better handled by searching k. Then, based on (18), p_{ij} can obtained as

$$p_{ij} = \frac{g_{i,k+1} - g_{i,j}}{k g_{i,k+1} - \sum_{j=1}^{k} g_{ij}}. \tag{19}$$

4) ALGORITHM

The procedure of the proposed AUFS method was summarized in Algorithm 1. It will stop when the objective function of equation (8) tends to a constant or the change is very close to zero. The most time consuming operation of Algorithm 1 is to solve the generalized eigenproblem in (11). The time complexity of this operation is $O(m^3)$ approximately, where m is the number of features.

The optimization procedure in Algorithm 1 will monotonically decrease the objective function in (7) in each iteration. Since the objective function has lower bounds (e.g., 0), the above iteration will converge. Besides, empirical results showed that the convergence of Algorithm 1 was fast. In the experiments, only several iterations (no more than 15 iterations) were needed to reach convergence.

Algorithm 1 The Proposed AUFS Method

Input: Gene expression data matrix $X \in \mathbb{R}^{n\times m}$; Parameters $k, c, q, \alpha, \beta, \gamma$; Number of selected genes d;
Output: d selected genes;

1: 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;
2: Initialize P^1 based on (19) by setting $g_{ij}^1 = \|x_i - x_j\|^2$;
3: Calculate $L_s^1 = H_n F^1 (F^1)^T H_n^T$ and $L_p^1 = D_p^1 - (P^1 + (P^1)^T)/2$;
4: Calculate $L^1 = \beta L_p^1 - L_s^1$;
5: Calculate W^1 by solving the generalized eigenproblem $(XL^1X^T + \alpha U^1)\tilde{w} = \lambda \tilde{X}\tilde{X}^T\tilde{w}$;
6: **repeat**
7: Calculate $M^t = -\tilde{X}^T W^t (W^t)^T \tilde{X}$;
8: Update $F_{ij}^{t+1} = F_{ij}^t \frac{(\gamma F^t)_{ij}}{(M^t F^t + \gamma F^t (F^t)^T F^t)_{ij}}$;
9: Update the diagonal matrix U^{t+1} with the i^{th} diagonal element as $U_{ii}^{t+1} = \frac{1}{2\|w_i^t\|_2}$;
10: Update P^{t+1} based on (19) and $g_{ij}^{t+1} = \|(W^t)^T x_i - (W^t)^T x_j\|^2$;
11: Calculate $L_s^{t+1} = H_n F^{t+1} (F^{t+1})^T H_n^T$ and $L_p^{t+1} = D_p^{t+1} - (P^{t+1} + (P^{t+1})^T)/2$;
12: Calculate $L^{t+1} = \beta L_p^{t+1} - L_s^{t+1}$;
13: Update W^{t+1} by solving the generalized eigenproblem $(XL^{t+1}X^T + +\alpha U^{t+1})\tilde{w} = \lambda \tilde{X}\tilde{X}^T\tilde{w}$;
14: t=t+1;
15: **until** Convergence
16: Sort each gene g_i based on $\|w_i\|_2$ in descending order. The top d ranked genes are selected.

III. MATERIALS

A. DATASET

We used a published gene expression dataset that has been studied in [16], which was primarily obtained from TCGA and generated by high-throughput techniques. This dataset contains two major subtypes of NSCLC: lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD). The objective is to distinguish these two subtypes of NSCLC. The samples without sufficient expression data have been excluded and finally 1013 samples were enrolled in the study, where 501 samples belonged to the LUSC subtype and 512 samples belonged to the LUAD subtype. Then, the differentially expressed (DE) genes in each subtype were identified with Edger [36]. All the DE genes were annotated based on the Ensembl database at http://asia.ensembl.org/. By selecting the genes (i.e., mRNAs and lncRNAs) that were differentially co-expressed in both LUAD and LUSC subtypes, 5469 genes (i.e., 3924 mRNAs and 1545 lncRNAs) were obtained for gene selection and classification.

B. INITIAL FILTERING OF GENES

To reduce the computational cost of the proposed feature selection method, we filtered out the genes that were less relevant to the two subtypes of NSCLC. Similar to the previous

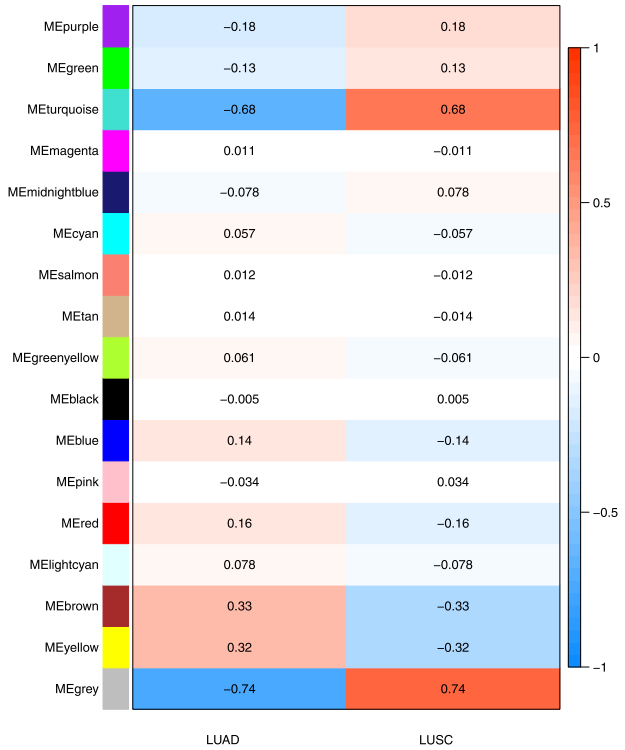

FIGURE 1. The correlation between modules and two subtypes of NSCLC.

study in [16], we constructed a dissimilarity matrix based on the topological overlap matrix (TOM), which reflected the gene-wise similarity. A hierarchical clustering tree with all the 5469 genes was then constructed based on the dissimilarity matrix, by which highly connected and co-expressed genes were clustered into some co-expression modules. The first principal component of each module, i.e., the eigengenes, was used to analyze the correlation between the modules and the subtype information. Only the genes in the most relevant modules were kept for further selection. The proposed AUFS method was then performed to select the candidate genes from each module and finally identified a small number of genes from the candidate genes as the informative genes for the NSCLC subtype classification.

IV. RESULTS

A. INITIAL FILTERING OF GENES

We performed module-based gene filtering before applying the proposed feature selection method, with the objective to reduce the computational cost and filter out the genes that were less relevant to the two subtypes of NSCLC. We used the methods in [16] to calculate the dissimilarity matrix, and performed hierarchical clustering to divide the 5469 genes into multiple modules. The relationship between the eigengenes of each module and the labels of two subtypes of NSCLC were analyzed by WGCNA [37]. The WGCNA function was implemented by 'WGCNA' (v1.63) [38]. Figure 1 shows the correlation between the modules and the two subtypes of NSCLC (i.e., LUAD and LUSC). 17 modules denoted by different colors were identified by hierarchical clustering. The modules that were most relevant to LUAD and LUSC are turquoise and grey. We used the genes in turquoise and grey modules for further selection using the proposed AUFS method. Finally, 960 genes in turquoise and 1882 genes in grey were reserved and would be used as input to AUFS.

B. EXPERIMENTAL SETTINGS

1) EVALUATION METRICS

The proposed method was evaluated based on accuracy, sensitivity, and specificity, which are defined as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{20}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{21}$$

and

$$Specificity = \frac{TN}{TN + FP}, \tag{22}$$

where TP is True Positives, i.e., the number of positive samples correctly classified as positive, TN is True Negatives, i.e., the number of negative samples correctly classified as negative, FP is False Positives, i.e., the number of negative samples incorrectly classified as positive, and FN is False Negatives, i.e., the number of positive samples incorrectly classified as negative [39].

2) PARAMETER SETTINGS

There are five parameters in the proposed method, i.e., the number of nearest neighbors k, the projected low-dimensions q, and three parameters in the objective function (8), i.e., α, β, and γ. In the experiments, k was set as 5, q was set as 2, and γ was set as 2^6. α and β were tuned over $\{1, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$. We utilized four classifiers, i.e., k-nearest neighbor (kNN) [40], Decision Tree (DT) [41], Support Vector Machine (SVM) [42], and Linear Discriminant Analysis (LDA) [43], for NSCLC subtype classification with the selected genes. We used the toolbox of Matlab to run the four classifiers with default settings. The training data and test data were set as 70% and 30% of the total samples, respectively. We created cross-validation partition for the samples using Matlab function "cvpartition". Unsupervised feature selection was performed to rank the genes. The classifiers were trained and tested using the top-ranked genes. All experiments were repeated 100 times and the mean results of test data were reported.

C. SELECTING CANDIDATE GENES FROM TWO MODULES

The proposed method was performed to select the candidate genes from the turquoise and grey modules, respectively.

1) COMPARISON WITH UNSUPERVISED FEATURE SELECTION METHODS

We compared the proposed AUFS method with several state-of-the-art unsupervised feature selection methods to select the candidate genes from the two modules. The compared methods includes LapScore [23], MCFS [26], NDFS [27], UDFS [30], and GLFS [24]. After initially selecting 960 genes in the turquoise module and 1882 genes in the

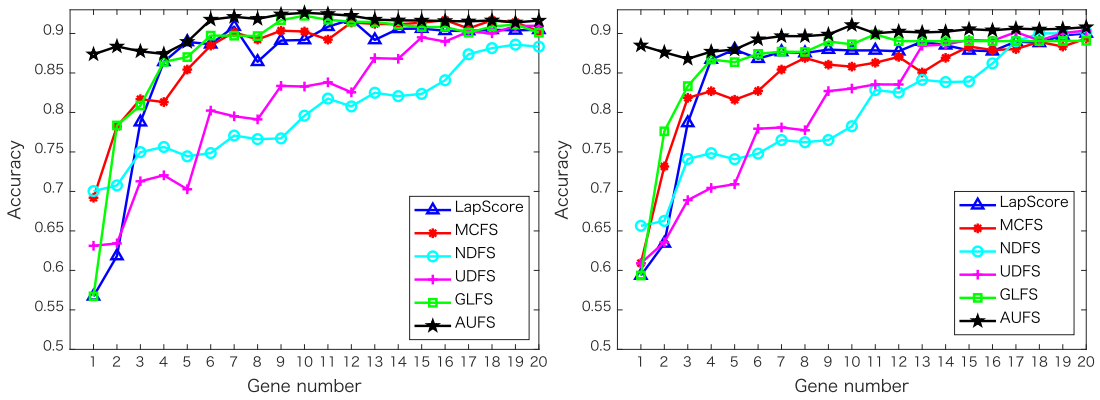

FIGURE 2. Comparison of different unsupervised feature selection methods in the turquoise module to select candidate genes by using (a) kNN and (b) SVM classifiers.

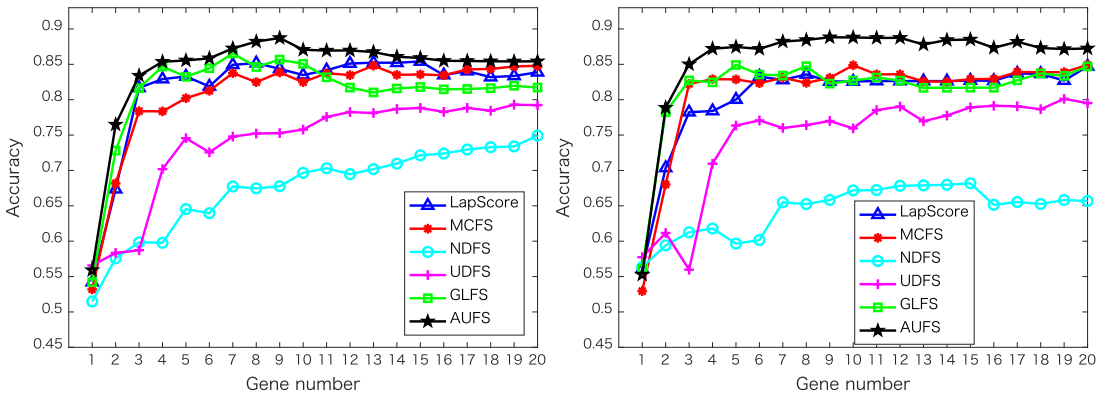

FIGURE 3. Comparison of different unsupervised feature selection methods in the grey module to select candidate genes by using (a) kNN and (b) SVM classifiers.

grey module, different unsupervised feature selection methods were performed to rank the genes in each module, respectively. The top-ranked genes were picked for training and testing. The number of picked genes varied from 1 to 20. The comparison results by the kNN and SVM classifiers in the turquoise and grey module are shown in Figures 2 and 3, respectively. The proposed AUFS method outperforms other compared methods. The curve of AUFS stays above the five other curves in each case. Note that in the turquoise module, the proposed method can maintain high accuracies when the number of selected genes is no more than three, which is a significant improvement compared to other methods. That is because the proposed method preserves the local structure based on only the relevant features, avoiding the adverse effects by the irrelevant features on the structure characterization for selecting the precise features.

2) SELECTED CANDIDATE GENES

Table 1 lists the top 10 genes ranked by the proposed method in the turquoise and grey modules, respectively. These genes are obtained by setting $\alpha = 10^6$, $\beta = 10^2$ in the turquoise module and $\alpha = 10^5$, $\beta = 10^2$ in the grey module. The 20 genes were selected as the candidate genes for further selection. The proposed method was performed on the

TABLE 1. Top-ranked genes obtained by AUFS in turquoise and grey modules.

Rank	Gene symbol (turquoise module)	Gene symbol (grey module)
1	*KRT6A*	*CXCL13*
2	*SFTPA2*	*PI3*
3	*SFTPA1*	*KRT14*
4	*GAPDH*	*S100A2*
5	*CEACAM6*	*SPRR1A*
6	*PERP*	*SPRR1B*
7	*AKR1B10*	*KRT16*
8	*PKP1*	*KRT6B*
9	*CSTA*	*S100A7*
10	*GPX2*	*SPP1*

20 candidate genes again to select the most informative genes as the signatures. We ranked the 20 candidate genes by the proposed method, and the top-ranked candidate genes were picked for training and testing. The number of picked candidate genes varied from 1 to 20. The classification results by the four classifiers (i.e., kNN, DT, SVM, and LDA) on the top-ranked candidate genes are shown in Figure 4. We can see from Figure 4 that all the four classifiers obtain the best results (i.e., highest accuracies) when the number of picked candidate genes equals to 17. These 17 candidate genes are considered to be the most relevant to the two subtypes of NSCLC. Therefore, we identified these 17 candidate genes

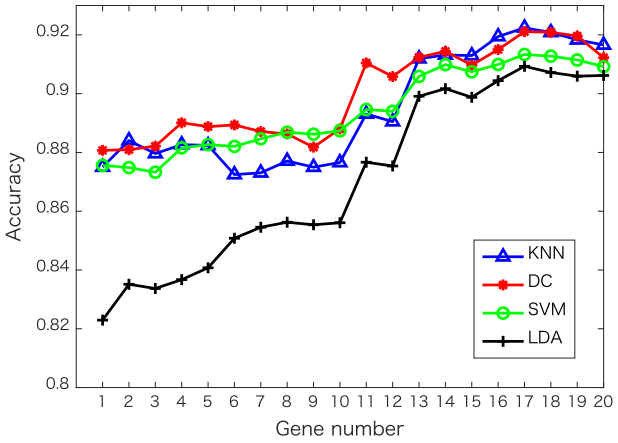

FIGURE 4. Classification results by four classifiers on top-ranked candidate genes.

TABLE 2. Gene signatures identified by AUFS.

Rank	Gene symbol	Module
1	*KRT6A*	turquoise
2	*SFTPA2*	turquoise
3	*SFTPA1*	turquoise
4	*KRT14*	grey
5	*PI3*	grey
6	*GAPDH*	turquoise
7	*S100A2*	grey
8	*KRT16*	grey
9	*SPRR1B*	grey
10	*SPRR1A*	grey
11	*CEACAM6*	turquoise
12	*KRT6B*	grey
13	*PERP*	turquoise
14	*S100A7*	grey
15	*AKR1B10*	turquoise
16	*PKP1*	turquoise
17	*CSTA*	turquoise

as the signatures and listed them in Table 2. Among them, 9 genes were from the turquoise module and 8 genes were from the grey module.

D. ANALYSIS OF SELECTED SIGNATURES

1) COMPARISON WITH OTHER NSCLC SUBTYPE CLASSIFICATION METHOD

Su *et al.* [16] proposed a gene selection method (called WGRFE) to select the gene signatures to distinguish LUSC from LUAD. They embedded the WGRFE and the standard RFE with SVM and RF to rank the genes, respectively. Their results demonstrated that RF-WGRFE achieved better performance than SVM-WGRFE. RF-WGRFE identified 13 gene signatures. We compared the proposed AUFS method using the 17 identified signatures with RF-WGRFE using the 13 identified signatures. Table 3 presents the comparison results by using four classifiers, i.e., kNN, DT, SVM, and LDA. Both AUFS and RF-WGRFE identified a small number of gene signatures. The genes KRT6A, KRT16, SPRR1B, KRT6B, PERP identified by AUFS were also identified by RF-WGRFE. We can see from Table 3, the proposed AUFS method performs better than RF-WGRFE in most cases.

TABLE 3. Comparison with Su *et al.*'s method.

Methods	# of genes	Classifiers	Accuracy	Sensitivity	Specificity
RF-WGRFE	13	kNN	91.2%	95.1%	88.0%
		DT	91.9%	92.2%	91.5%
		SVM	91.3%	96.6%	91.2%
		LDA	89.8%	96.3%	83.2%
AUFS	17	kNN	92.2%	96.1%	88.3%
		DT	92.1%	94.6%	90.1%
		SVM	91.3%	95.8%	86.8%
		LDA	91.0%	97.5%	84.3%

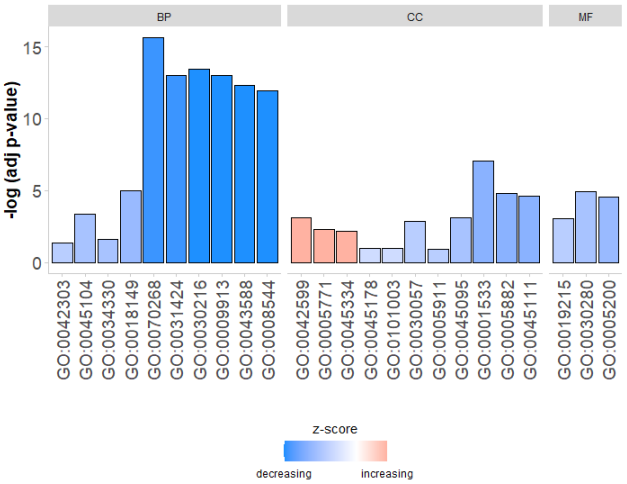

FIGURE 5. The significantly enriched GO terms. Each bar represents a GO term.

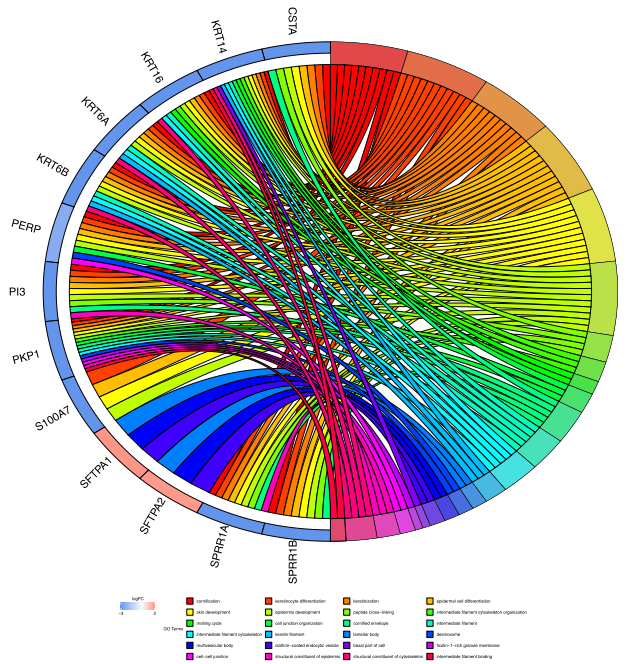

FIGURE 6. The top significantly enriched GO terms. The left half of the circle is the signature and each color corresponds to a GO term in right half of the circle.

2) FUNCTIONAL ENRICHMENT ANALYSIS

We used Gene Ontology (GO) [44] to analyze the biological meaning of the gene signatures identified by the proposed AUFS method. The GO enrichment analysis was

implemented by using the packages in [45]. Specific GO including biological process (BP), cellular component (CC) and molecular function (MF) were investigated. We analyzed the enrichment of the 17 identified gene signatures in Table 2. The threshold for significant enrichment was set as p-value=0.05. Figure 5 shows the significantly enriched GO terms. Each bar represents a GO term and a higher bar means a higher degree of enrichment. We can see from figure 5 that 24 terms were obtained, which includes 10 BP terms, 11 CC terms, and 3 MF terms. 13 out of the 17 signatures were verified to present biological meaning and the top significantly enriched GO terms are shown in figure 6. Compared to the 13 gene signatures identified in [16], the 17 identified gene signatures obtained lower p-values (as shown in figure 5) and more signatures were verified to present biological meaning (as shown in figure 6). We also note that cornification and keratinocyte differentiation are the most significant terms. The term keratinization has been reported to be associated with LUSC in some previous studies [46].

V. CONCLUSION

The increasing availability of genome-wide gene expression data has facilitated the identification of gene signatures for precise NSCLC subtypes classification. Most existing feature selection methods applied to signature identification are supervised, which need class labels or related gene information. To utilize the useful information hidden in data, we proposed a novel unsupervised feature selection method to identify the most discriminative gene signatures for the NSCLC subtype classification. The proposed method incorporated discriminant analysis, adaptive structure preservation, and $l_{2,1}$-norm sparse regression into a joint learning framework to select the informative genes. We developed an effective algorithm to solve the optimization problem in the proposed method. Furthermore, we perform module-based gene filtering before feature selection to reduce the computational cost and filter out the genes that were less relevant to the subtypes of NSCLC. We evaluate the proposed method on a published gene expression dataset of NSCLC from TCGA, which contained two major subtypes LUAD and LUSC. The experimental results demonstrate that the proposed method identified a small number of gene signatures for accurate subtype classification: i.e., distinguishing LUSC from LUAD. Enrichment analysis of the selected gene signatures was also performed by summarizing the key biological processes. The results show that cornification and keratinocyte differentiation are the most significant GO terms.

Although the proposed method is applied to identify the gene signatures for the NSCLC subtype classification based on gene expression data, the proposed method is generally applicable to other types of biological data and other types of tumors. In this study, we mainly focus on distinguishing the two subtypes LUSC and LUAD of NSCLC. In future work, we will apply the proposed method to classify more other subtypes of NSCLC and also other types of biological data.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[2] Y. Wang, Y. Wang, Y. Wang, and Y. Zhang, "Identification of prognostic signature of non–small cell lung cancer based on TCGA methylation data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Dec. 2020.

[3] R. He and S. Zuo, "A robust 8-gene prognostic signature for early-stage non-small cell lung cancer," *Frontiers Oncol.*, vol. 9, p. 693, Jul. 2019.

[4] F. Janku, D. J. Stewart, and R. Kurzrock, "Targeted therapy in non-small-cell lung cancer—Is it becoming a reality?" *Nature Rev. Clin. Oncol.*, vol. 7, no. 7, p. 401, 2010.

[5] L. Girard, J. Rodriguez-Canales, C. Behrens, D. M. Thompson, I. W. Botros, H. Tang, Y. Xie, N. Rekhtman, W. D. Travis, I. I. Wistuba, J. D. Minna, and A. F. Gazdar, "An expression signature as an aid to the histologic classification of non-small cell lung cancer," *Clin. Cancer Res.*, vol. 22, no. 19, pp. 4880–4889, Oct. 2016.

[6] Y.-K. Zhang, W.-Y. Zhu, J.-Y. He, D.-D. Chen, Y.-Y. Huang, H.-B. Le, and X.-G. Liu, "MiRNAs expression profiling to distinguish lung squamous-cell carcinoma from adenocarcinoma subtypes," *J. Cancer Res. Clin. Oncol.*, vol. 138, no. 10, pp. 1641–1650, Oct. 2012.

[7] F.-Q. Nie, Q. Zhu, T.-P. Xu, Y.-F. Zou, M. Xie, M. Sun, R. Xia, and K.-H. Lu, "Long non-coding RNA MVIH indicates a poor prognosis for non-small cell lung cancer and promotes cell proliferation and invasion," *Tumor Biol.*, vol. 35, no. 8, pp. 7587–7594, Aug. 2014.

[8] X. Wang and O. Gotoh, "A robust gene selection method for microarray-based cancer classification," *Cancer Informat.*, vol. 9, p. S3794, Jan. 2010.

[9] X. Qiang, C. Zhou, X. Ye, P.-F. Du, R. Su, and L. Wei, "CPPred-FL: A sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning," *Briefings Bioinf.*, vol. 21, no. 1, pp. 11–23, 2020.

[10] X. Ye, H. Li, T. Sakurai, and P.-W. Shueng, "Ensemble feature learning to identify risk factors for predicting secondary cancer," *Int. J. Med. Sci.*, vol. 16, no. 7, pp. 949–959, 2019.

[11] L. Wei, J. Hu, F. Li, J. Song, R. Su, and Q. Zou, "Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms," *Briefings Bioinf.*, vol. 21, no. 1, pp. 106–119, 2020.

[12] R. Su, H. Wu, X. Liu, and L. Wei, "Predicting drug-induced hepatotoxicity based on biological feature maps and diverse classification strategies," *Briefings Bioinf.*, Dec. 2019, doi: 10.1093/bib/bbz165.

[13] P. A. Mundra and J. C. Rajapakse, "Support vector based T-score for gene ranking," in *Proc. IAPR Int. Conf. Pattern Recognit. Bioinf.* Melbourne, VIC, Australia: Springer, 2008, pp. 144–153.

[14] O. Reyes, C. Morell, and S. Ventura, "Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context," *Neurocomputing*, vol. 161, pp. 168–182, Aug. 2015.

[15] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.

[16] R. Su, J. Zhang, X. Liu, and L. Wei, "Identification of expression signatures for non-small-cell lung carcinoma subtype classification," *Bioinformatics*, vol. 36, no. 2, pp. 339–346, 2020.

[17] M. Johannes, J. C. Brase, H. Fröhlich, S. Gade, M. Gehrmann, M. Fälth, H. Sültmann, and T. Beißbarth, "Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients," *Bioinformatics*, vol. 26, no. 17, pp. 2136–2144, Sep. 2010.

[18] R. Su, X. Liu, and L. Wei, "MinE-RFE: Determine the optimal subset from RFE by minimizing the subset-accuracy–defined energy," *Briefings Bioinf.*, vol. 21, no. 2, pp. 687–698, Mar. 2020.

[19] C. Winter, G. Kristiansen, S. Kersting, J. Roy, D. Aust, T. Knösel, P. Rümmele, B. Jahnke, F. Rückert, and M. Niedergethmann, "Google goes cancer: Improving outcome prediction for cancer patients by network-based ranking of marker genes," *PLoS Comput. Biol.*, vol. 8, no. 5, May 2012, Art. no. e1002511.

[20] X. Ye and T. Sakurai, "Robust similarity measure for spectral clustering based on shared neighbors," *ETRI J.*, vol. 38, no. 3, pp. 540–550, 2016.

[21] X. Ye, H. Li, A. Imakura, and T. Sakurai, "Distributed collaborative feature selection based on intermediate representation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 563–568.

[22] X. Ye and T. Sakurai, "Feature selection via embedded learning based on tangent space alignment for microarray data," *J. Comput. Sci. Eng.*, vol. 11, no. 4, pp. 121–129, Dec. 2017.
[23] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 507–514.
[24] X. Ye, K. Ji, and T. Sakurai, "Global discriminant analysis for unsupervised feature selection with local structure preservation," in *Proc. 29th Int. Florida Artif. Intell. Res. Soc. Conf.*, 2016, pp. 454–459.
[25] X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu, "Global and local structure preservation for feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 6, pp. 1083–1095, Jun. 2014.
[26] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 333–342.
[27] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 1026–1032.
[28] X. Ye, K. Ji, and T. Sakurai, "Unsupervised feature selection with correlation and individuality analysis," *Int. J. Mach. Learn. Comput.*, vol. 6, no. 1, pp. 36–41, 2016.
[29] J. Tang, X. Hu, H. Gao, and H. Liu, "Discriminant analysis for unsupervised feature selection," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2014, pp. 9–17.
[30] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou, "$\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1589–1594.
[31] X. Ye and T. Sakurai, "Unsupervised feature selection for microarray gene expression data based on discriminative structure learning," *J. Universal Comput. Sci.*, vol. 24, no. 6, pp. 725–741, 2018.
[32] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 977–986.
[33] X. Ye and T. Sakurai, "Spectral clustering with adaptive similarity measure in kernel space," *Intell. Data Anal.*, vol. 22, no. 4, pp. 751–765, Jun. 2018.
[34] H. Tao, C. Hou, F. Nie, Y. Jiao, and D. Yi, "Effective discriminative feature selection with nontrivial solution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 796–808, Apr. 2016.
[35] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[36] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "EdgeR: A bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010.
[37] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Stat. Appl. Genet. Mol. Biol.*, vol. 4, no. 1, 2005.
[38] P. Langfelder and S. Horvath, "WGCNA: An R package for weighted correlation network analysis," *BMC Bioinf.*, vol. 9, no. 1, p. 559, Dec. 2008.
[39] X. Ye, H. Li, A. Imakura, and T. Sakurai, "An oversampling framework for imbalanced classification based on Laplacian eigenmaps," *Neurocomputing*, vol. 399, pp. 107–116, Jul. 2020.
[40] M. J. Islam, Q. M. Jonathan Wu, M. Ahmadi, and M. A. Sid-Ahmed, "Investigating the performance of Naive-Bayes classifiers and K-nearest neighbor classifiers," in *Proc. Int. Conf. Converg. Inf. Technol. (ICCIT)*, Nov. 2007, pp. 1541–1546.
[41] W.-Y. Loh, "Classification and regression trees," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 14–23, 2011.
[42] A. M. Andrew, "An introduction to support vector machines and other kernel-based learning methods by Nello Christianini and John Shawe-Taylor, Cambridge University press, Cambridge, 2000, xiii+189 pp., isbn 0-521-78019-5 (hbk,£27.50)," *Robotica*, vol. 18, no. 6, pp. 687–689, 2000.
[43] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936.
[44] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: Tool for the unification of biology," *Nature Genet.*, vol. 25, no. 1, pp. 25–29, May 2000.
[45] G. Yu, L.-G. Wang, G.-R. Yan, and Q.-Y. He, "DOSE: An R/bioconductor package for disease ontology semantic and enrichment analysis," *Bioinformatics*, vol. 31, no. 4, pp. 608–609, Feb. 2015.
[46] H. J. Park, Y.-J. Cha, S. H. Kim, A. Kim, E. Y. Kim, and Y. S. Chang, "Keratinization of lung squamous cell carcinoma is associated with poor clinical outcome," *Tuberculosis Respiratory Diseases*, vol. 80, no. 2, pp. 179–186, 2017.

XIUCAI YE (Member, IEEE) received the Ph.D. degree in computer science from the University of Tsukuba, Tsukuba, Japan, in 2014. She is currently an Assistant Professor with the Department of Computer Science, and Center for Artificial Intelligence Research (C-AIR), University of Tsukuba. Her current research interests include feature selection, clustering, machine learning and its application fields, and bioinformatics.

WEIHANG ZHANG is currently pursuing the master's degree with the Department of Computer Science, University of Tsukuba, Japan. His current research interests include data analysis and machine learning.

TETSUYA SAKURAI (Member, IEEE) received the Ph.D. degree in computer engineering from Nagoya University, in 1992. He is currently a Professor of the Department of Computer Science, and the Director of the Center for Artificial Intelligence Research (C-AIR), University of Tsukuba. He is also a Visiting Professor with the Open University of Japan and also a Visiting Researcher of the Advanced Institute of Computational Science, RIKEN. His research interests include high performance algorithms for large-scale simulations, data and image analysis, and deep neural network computations. He is a member of the Japan Society for Industrial and Applied Mathematics (JSIAM), the Mathematical Society of Japan (MSJ), the Information Processing Society of Japan (IPSJ), and the Society for Industrial and Applied Mathematics (SIAM).