Hindawi
BioMed Research International
Volume 2021, Article ID 6490118, 11 pages
https://doi.org/10.1155/2021/6490118

Research Article

Gene Correlation Guided Gene Selection for Microarray Data Classification

Dong Yang[1] **and Xuchang Zhu**[2]

[1]*Department of Colorectal Surgery, Tianjin Union Medical Center, Tianjin 300121, China*
[2]*Department of Gastrointestinal Surgery, Lianshui People's Hospital Affiliated to Kangda College of Nanjing Medical University, Huai'an 223300, China*

Correspondence should be addressed to Xuchang Zhu; 0402046@wust.edu.cn

Received 18 June 2021; Accepted 9 August 2021; Published 16 August 2021

Academic Editor: Chang Tang

The microarray cancer data obtained by DNA microarray technology play an important role for cancer prevention, diagnosis, and treatment. However, predicting the different types of tumors is a challenging task since the sample size in microarray data is often small but the dimensionality is very high. Gene selection, which is an effective means, is aimed at mitigating the curse of dimensionality problem and can boost the classification accuracy of microarray data. However, many of previous gene selection methods focus on model design, but neglect the correlation between different genes. In this paper, we introduce a novel unsupervised gene selection method by taking the gene correlation into consideration, named gene correlation guided gene selection (G^3CS). Specifically, we calculate the covariance of different gene dimension pairs and embed it into our unsupervised gene selection model to regularize the gene selection coefficient matrix. In such a manner, redundant genes can be effectively excluded. In addition, we utilize a matrix factorization term to exploit the cluster structure of original microarray data to assist the learning process. We design an iterative updating algorithm with convergence guarantee to solve the resultant optimization problem. Experimental results on six publicly available microarray datasets are conducted to validate the efficacy of our proposed method.

1. Introduction

During cell division and growth, abnormal changes often happen to genes, which results in varying cancers. With the rapid development of kinds of biomedical technologies [1], DNA microarray comes into being and lots of microarray data can be obtained for cancer prevention, diagnosis, and treatment [2–12]. For various microarray data, classifying the different types of tumors is an important task, but challenging due to the high dimensionality and small numbers of samples [13–15] since the small number of data samples with large number of genes can easily result in the "curse of dimensionality" and overfitting problems of data processing and learning models. When the dimension of samples is too high, the distance between any two samples is very inaccurate. Therefore, the classification task for this kind of data is often challenging. However, it has been verified by some existing biological experiments that only a very small proportion of genes contribute significantly to biological process and disease indication. Directly processing original high dimensional microarray data not only degenerates the classification performance but also brings extra computation burden of hardware. Therefore, it is necessary to select a subset of discriminative genes from high-dimensional microarray data to serve subsequent tasks [16–25]. If we treat each gene as a feature dimension in microarray data, gene selection is similar to the feature selection task in machine learning and data mining community [26–37]. In fact, many feature selection methods can be used well for gene selection. Therefore, mathematical gene selection methods can be also grouped into three classes, i.e., filter methods, wrapper methods, and embedded methods.

Filter methods often measure the importance of different genes in a straight-forward manner based on certain criteria such as t-test [38, 39], Z-score [40], signal-to-noise ratio (SNR) [41], Laplacian score [42], mutual information [43], and information gain [44]. In [41], Golub et al. firstly used the SNR function to evaluate the weights of the genes. Many traditional feature selection methods such as ReliefF [45] and MRMR [46] are combined and used for gene selection [47]. Since filter methods only depend on the intrinsic properties of original data [48], a good ranking criterion function is very important.

As to wrapper methods, varying classification algorithms are often used as a fitness evaluation to determine the subset of genes and the selected genes can in turn enhance the classification performance [2, 49–56]. In general, wrapper methods can obtain better results than filter methods, but bring more expensive computational cost. A lot of evolutionary algorithms such as genetic algorithm (GA), differential evolution (DE), ant colony optimization (ACO), and simulated annealing are commonly used as wrapper methods for gene selection [57, 58].

For embedded methods, the geometric structure and intrinsic property of data are exploited to construct gene selection models. Among this kind of methods, some mathematical regularization terms with specific physic meanings such as representative and sparse characters are commonly used assumptions. Typical models include self-representation [32, 33, 59–62], low-rank representation [63, 64], and matrix factorization [65–67]. Based on these basic models, many variants have been proposed, such as Laplacian graph regularized low-rank representation [63]. Considering the robustness to outliers, Wang et al. [66] proposed a robust $l_{2,1}$-norm regularized characteristic gene selection method. In [68], Guo et al. proposed to identify the disease-associated genes by utilizing ensemble consensus-guided unsupervised feature selection method. In an unsupervised manner, the major priori information can be used is the intrinsic local geometric structure of data. Therefore, embedded methods that use this priori information can achieve good performance for various of microarray data and obtain more and more attention.

Although there are lots of computational methods proposed for gene selection and achieve great success, most of them focus on the relation of data samples while the correlation between different genes is ignored. The expression values of different genes should be interrelated for a certain microarray data matrix. Therefore, we propose to calculate the correlation of gene pairs to regularize the gene selection model, which is named as named gene correlation guided gene selection (G^{3}CS). In detail, in order to exclude redundant genes, the covariance of different gene dimension pairs is calculated and embedded into our unsupervised gene selection model to regularize the gene selection coefficient matrix. In addition, we utilize a matrix factorization model which can capture the cluster structure of original data to assist the learning process. We design an iterative updating algorithm to solve the resultant problem. Finally, experimental results on six publicly available real microarray datasets are conducted to demonstrate that the proposed G^{3}CS can steadily perform better than other state-of-the-art computational gene selection methods in terms of microarray data classification. In Figure 1, we give a brief flowchart of our proposed G^{3}CS model.

2. Related Work

In this section, we introduce some gene selection works that are most related to our proposed method. Before that, we firstly present some notations will be used in the following sections. Throughout this paper, matrices and vectors are denoted as boldface capital letters boldface lower case letters, respectively. Given an matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\mathbf{X}_{ij}$ represents its (i, j)-th element, $\mathbf{x}^i$ and $\mathbf{x}_j$ denotes its i-th row and j-th column, respectively. $\mathbf{X}^T$ is the transpose of $\mathbf{X}$. If $\mathbf{X}$ is square, $Tr(\mathbf{X})$ is the trace of $\mathbf{X}$. $\mathbf{I}_k$ denotes an identity matrix with size $k \times k$. $\mathbf{1}$ is a vector with all elements are 1. $\|\mathbf{X}\|_{2,1} = \sum_{i=1}^{m} \|\mathbf{x}^i\| = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} \mathbf{X}_{ij}^2}$ denotes the $l_{2,1}$-norm of matrix $\mathbf{X}$, which is used to constrain the row sparsity of $\mathbf{X}$. $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{X}_{ij}^2}$ is the well-known Frobenius norm of $\mathbf{X}$.

Since our proposed G^{3}CS belongs to the embedded method, we give a brief review about some related embedded methods.

2.1. GRSL-GS. In [20], Tang et al. proposed a manifold regularized subspace learning model for gene selection, in which the model projects original high dimensional microarray data into a lower-dimensional subspace, then original genes are constrained to be well represented by the selected gene subset. In order to capture the local manifold structure of original data, a Laplacian graph regularization term is imposed on the low-dimensional data space. Finally, the learned projection matrix can be regarded as an important indicator of different genes. Specifically, the mathematical model of GRSL-GS can be formulated as follows:

$$\begin{aligned} &\arg\min_{\mathbf{C},\mathbf{P}} \|\mathbf{X} - \mathbf{XPC}\|_F^2 + \lambda Tr\left(\mathbf{P}^T\mathbf{X}^T\mathbf{LXP}\right) \\ &\text{s.t.}\mathbf{P} \geq 0, \mathbf{C} \geq 0, \mathbf{P}^T\mathbf{P} = \mathbf{I}, \end{aligned} \tag{1}$$

where $\mathbf{P}$ denotes the projection matrix, $\mathbf{C}$ represents the data reconstruction coefficient matrix, and $\mathbf{L}$ is the Laplacian matrix calculated from original data. λ is a hyperparameter that balances the two regularization terms. The first term in Eq. (1) constraints that original microarray data can be reconstructed from the projected lower-dimensional gene dictionary, and the second term is the graph Laplacian regularization term used to preserve the intrinsic local manifold structure of original data samples. Although GRSL-GS captures the local structure information, it does not exploit the gene correlation.

2.2. AHEDL. Considering that the graph Laplacian in GRSL-GS can only capture pairwise sample relationship, Zheng et al. [22] introduced a computational gene selection model via adaptive hypergraph embedded dictionary learning (AHEDL). Similar to GRSL-GS, AHEDL also learns a

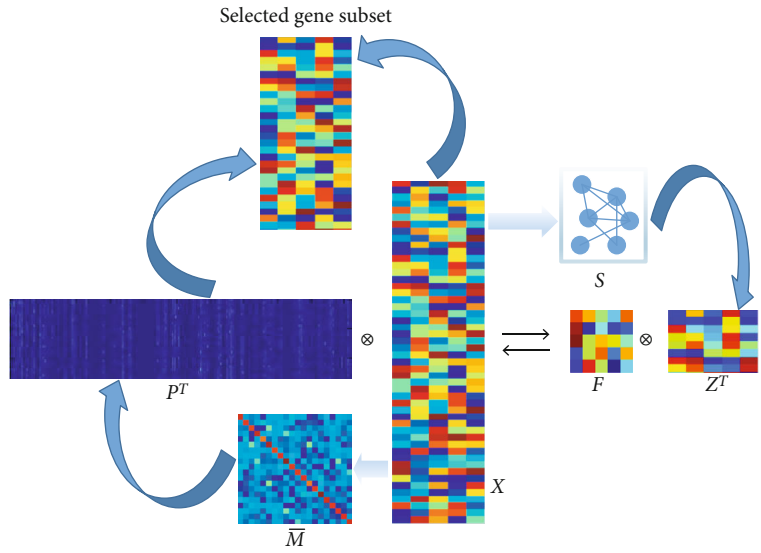

Figure 1: Brief flowchart of our proposed G^{3}CS model.

dictionary from original high dimensional microarray data, and the learned dictionary is then used to represent original data by a reconstruction coefficient matrix. The difference of dictionary learning between GRSL-GS and AHEDL is that GRSL-GS uses projection process but AHEDL directly utilizes traditional dictionary learning model. The $l_{2,1}$ -norm is imposed on the coefficient matrix for selecting discriminate genes.

In addition, the hypergraph is also learned in an adaptive manner. In a nutshell, AHEDL can be formulated as follows:

$$\min_{\mathbf{D},\mathbf{C},\mathbf{H}_e,\mathbf{H}_v,\mathbf{W}} \|\mathbf{X}-\mathbf{DC}\|_F^2 + \alpha Tr\left(\mathbf{CLC}^T\right) + \beta\|\mathbf{C}\|_{2,1} + \gamma Tr\left(\mathbf{W}^T\mathbf{W}\right)$$
$$\text{s.t.}\|\mathbf{d}_i\|^2 \le 1, \mathbf{w}^T\mathbf{1} = \mathbf{1}, w(e_i) > 0, \tag{2}$$

As can be seen from Eq. (2), AHEDL integrates adaptive hypergraph learning, dictionary learning, and gene selection into a uniform framework. The dictionary matrix $\mathbf{D}$, representation coefficient matrix $\mathbf{C}$ and hypergraph $\mathbf{W}$ can constrain each other during the optimization process to obtain their optimums. Since $\mathbf{D}$ can be regarded as the new representation of $\mathbf{X}$ in the dictionary space, the row sparsity imposed on $\mathbf{C}$ by using the $l_{2,1}$-norm can be used to measure the importance of gene dimensions in the learned dictionary space.

3. Proposed Method

Given a microarray data $\mathbf{X} \in \mathbf{R}^{m\times n}$, which contains n data samples with m different genes. Gene selection aims to select a gene subset that contains only a small number of genes for subsequent tasks. Without sample label information, we should exploit the intrinsic structure of data as much as possible. In this work, we deploy traditional regression model as the basic architecture to formulate G^{3}CS, which can be represented as follows:

$$\min_{\mathbf{P}} \left\|\mathbf{P}^T\mathbf{X} - \mathbf{C}\right\|_F^2 + \alpha\|\mathbf{P}\|_{2,1}, \tag{3}$$

where $\mathbf{P} \in \mathrm{R}^{m\times c}$ is a projection matrix that projects original data into label space $\mathbf{C} = [\mathbf{c}_1,\cdots,\mathbf{c}_n] \in \{0,1\}^{c\times n}$, where $\mathbf{c}_i \in \{0,1\}^c$ is the cluster indicator vector corresponding to $\mathbf{x}_i$. In order to measure the importance of different genes, we impose the $l_{2,1}$-norm on $\mathbf{P}$ to constrain that important genes contribute more during the projection process. In machine learning and data mining community, matrix factorization of target matrix $\mathbf{C}$ often shows remarkable performance [67, 69]. In our G^{3}CS model, we also decompose $\mathbf{C}$ into two components, i.e., $\mathbf{F} \in R^{c\times c}$ and $\mathbf{Z} \in R^{n\times c}$. As a result, Eq. (3) can be rewritten as following form with appropriate constraints:

$$\min_{\mathbf{P},\mathbf{F},\mathbf{Z}} \left\|\mathbf{P}^T\mathbf{X} - \mathbf{F}\mathbf{Z}^T\right\|_F^2 + \alpha\|\mathbf{P}\|_{2,1}$$
$$\text{s.t.}\mathbf{F}^T\mathbf{F} = \mathbf{I}, \mathbf{Z}^T\mathbf{Z} = I, \mathbf{Z} \ge 0, \tag{4}$$

where $\mathbf{F}^T\mathbf{F} = \mathbf{I}$ constrain each column of $\mathbf{B}$ to be independent with each other. $\mathbf{Z}^T\mathbf{Z} = \mathbf{I}$ is a relaxation constraint that makes each row of $\mathbf{Z}$ to have only one nonzero element. The constraints in Eq. (8) make the model to conduct orthogonal clustering which works well for unsupervised feature selection [70].However, by minimizing Eq. (8) directly for gene selection neglects the gene correlation information which is important in biomedical process. In this work, we embed the gene correlation information into G^{3}CS. It is well known that in probability theory and statistics, a covariance matrix is a square matrix giving the covariance between each pair of elements of a given random vector. In this work, we use covariance to calculate the correlation of different gene pairs, then, we can get a symmetric semipositive definite covariance

matrix $\mathbf{M}$. The i,j-th entry of covariance matrix $\mathbf{M}$ can be calculated as follows:

$$\mathbf{M}(i,j) = \frac{\sum_{i=1}^{m}\left(\mathbf{x}^i - \bar{\mathbf{x}}\right)\left(\mathbf{x}^i - \bar{\mathbf{x}}\right)}{m-1}, \tag{5}$$

where $\bar{\mathbf{x}}$ is the gene average vector, which is calculated as follows:

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{m}\mathbf{x}^i}{m}. \tag{6}$$

However, the diagonal elements in $\mathbf{M}$ only reflect the relationship between a gene dimension and itself, which makes no sense in our model. Therefore, we adjust $\mathbf{M}$ to get a new correlation matrix $\bar{\mathbf{M}}$ by the following equation:

$$\bar{\mathbf{M}}(i,j) = \begin{cases} \mathbf{M}(i,j) & \text{if } i = j, \\ \sum_{k \neq i}\mathbf{M}(i,k) & \text{if } i \neq j. \end{cases} \tag{7}$$

In such a manner, $\bar{\mathbf{M}}(i,j)$ represents the correlation between the i-th gene dimension with all other gene dimensions. Then, $\bar{\mathbf{M}}(i,j)$ can be embedded into Eq. (8) to emphasize the independence of selected gene dimensions from the perspective of data information. Therefore, we have

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{F},\mathbf{Z}} \left\|\mathbf{P}^T\mathbf{X} - \mathbf{F}\mathbf{Z}^T\right\|_F^2 + \alpha\|\mathbf{P}\|_{2,1} + \beta Tr\left(\mathbf{P}^T\bar{\mathbf{M}}\mathbf{P}\right), \\ &\text{s.t.}\mathbf{F}^T\mathbf{F} = \mathbf{I}, \mathbf{Z}^T\mathbf{Z} = \mathbf{I}, \mathbf{Z} \geq 0. \end{aligned} \tag{8}$$

In addition, the local geometric structure information of original data should be preserved as much as possible in the learned new space $\mathbf{Z}$. By using the Gaussian kernel function, we can get a similarity matrix from original data by the following equation:

$$\mathbf{S}_{ij} = \begin{cases} \exp\left(\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{-2t^2}\right), & \mathbf{x}_i \in \mathcal{N}_k\left(\mathbf{x}_j\right) \text{ or } \mathbf{x}_j \in \mathcal{N}_k\left(\mathbf{x}_i\right); \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where $\mathcal{N}_k(\mathbf{x}_i)$ represents the set of k nearest neighbors of $\mathbf{x}_i$, and t is a width parameter. k and t are set to 5 and 0.5, respectively, in our experiments. In our G^{3}CS model, we require that if two data samples are closed to each other in original space, their cluster indicator vectors in new space $\mathbf{Z}$ should also be close. This constraint can be formulated by using the following form:

$$\min_{\mathbf{z}} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\mathbf{z}^i - \mathbf{z}^j\right\|_2^2\mathbf{S}(i,j) = \min_{\mathbf{z}} Tr\left(\mathbf{Z}^T\mathbf{L}\mathbf{Z}\right), \tag{10}$$

where $\mathbf{L}$ is the Laplacian matrix corresponding to $\mathbf{S}$. Finally, we get the mathematical formulation of our G^{3}CS model as follows:

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{F},\mathbf{Z}} \left\|\mathbf{P}^T\mathbf{X} - \mathbf{F}\mathbf{Z}^T\right\|_F^2 + \alpha\|\mathbf{P}\|_{2,1} + \beta Tr\left(\mathbf{P}^T\bar{\mathbf{M}}\mathbf{P}\right) + \gamma Tr\left(\mathbf{Z}^T\mathbf{L}\mathbf{Z}\right) \\ &\text{s.t.}\mathbf{F}^T\mathbf{F} = \mathbf{I}, \mathbf{Z}^T\mathbf{Z} = \mathbf{I}, \mathbf{Z} \geq 0. \end{aligned} \tag{11}$$

where α, β, and γ are three hyperparameters to balance different regularization terms. In summary, Eq. (1) integrates regression, matrix factorization, gene correlation, and data local structure exploitation into a unified framework. The gene correlation regularizes the model to exclude redundant gene dimensions.

4. Optimization Algorithm

There are three variables in Eq. (1) that need to be optimized; we cannot obtain a close-form solution simultaneously for all of them. Therefore, we design an algorithm to iteratively update these variables. For each time, we update a variable by fixing other ones.

4.1. Optimize P. When other variables are fixed, solving $\mathbf{P}$ is equal to the following problem:

$$\min_{\mathbf{P}} \left\|\mathbf{P}^T\mathbf{X} - \mathbf{F}\mathbf{Z}^T\right\|_F^2 + \alpha\|\mathbf{P}\|_{2,1} + \beta Tr\left(\mathbf{P}^T\bar{\mathbf{M}}\mathbf{P}\right). \tag{12}$$

By taking the derivative of Eq. (12) with respect to $\mathbf{P}$ and setting it to zero, we have

$$2\mathbf{X}\mathbf{X}^T\mathbf{P} - 2\mathbf{X}\mathbf{Z}\mathbf{F}^T + 2\alpha\mathbf{G}\mathbf{P} + 2\beta\bar{\mathbf{M}}\mathbf{P} = 0, \tag{13}$$

Then, we have the closed-form solution of $\mathbf{P}$ as follows:

$$\mathbf{P} = \left(\mathbf{X}\mathbf{X}^T + \alpha\mathbf{G} + \beta\bar{\mathbf{M}}\right)^{-1}\mathbf{X}\mathbf{Z}\mathbf{F}^T, \tag{14}$$

where $\mathbf{G}$ is a diagonal matrix with $\mathbf{G}_{ii} = 1/2\|\mathrm{P}^i\|_2$. At each iteration, $\mathbf{G}$ and $\mathbf{P}$ can be updated alternatively.

4.2. Optimize F. When fixing other variables, the optimization problem is equal to the following equation:

$$\min_{\mathbf{F}} \left\|\mathbf{P}^T\mathbf{X} - \mathbf{F}\mathbf{Z}^T\right\|_F^2 s.t.\mathbf{F}^T\mathbf{F} = \mathbf{I}. \tag{15}$$

By adding a constant matrix $\mathbf{Z}$ into the F-norm, Eq. (15) is equal to

$$\min_{\mathbf{F}} \left\|\left(\mathbf{P}^T\mathbf{X} - \mathbf{F}\mathbf{Z}^T\right)\mathbf{Z}\right\|_F^2 s.t.\mathbf{F}^T\mathbf{F} = \mathbf{I}. \tag{16}$$

Since $\mathbf{Z}$ is an orthogonal matrix, then, we have

$$\min_{\mathbf{F}} \|\boldsymbol{W} - \boldsymbol{F}\|_F^2 s.t.\mathbf{F}^T\mathbf{F} = \mathbf{I}. \tag{17}$$

where $\mathbf{W} = \mathbf{P}^T\mathbf{X}\mathbf{Z}$. In order to ensure the orthogonal constraint of $\mathbf{F}$, we add a large positive constant ρ and the optimization problem can be converted to

Input: Microarray data matrix $\mathbf{X}$, parameters: α, β and γ. A small constant ε =0.0000001.
Initialize: $\mathbf{M}$, $\mathbf{S}$, $\mathbf{F}$, and $\mathbf{Z}$.
While not converged do
1. Update $\mathbf{P}$ via Eq.(14);
2. Update $\mathbf{F}$ via Eq.(20);
3. Update $\mathbf{Z}$ by solving Eq.(24);
6. Check convergence condition: $|obj^{t-1} - obj^t|/obj^t < \varepsilon$.
End while
Output: $\mathbf{P}$.
Gene selection: Sort the l_2-norm of the rows of $\mathbf{P}$ in decent order and select the largest K values. The gene dimension indexes with the the largest K values are selected to form the gene subset.

Algorithm 1: Optimization algorithm of the proposed G^{3}CS model.

$$\min_{\mathbf{F}} \frac{1}{2}\|\mathbf{W}-\mathbf{F}\|_F^2 + \frac{\rho}{4}\left\|\mathbf{F}^T\mathbf{F}-\mathbf{I}\right\|_F^2. \tag{18}$$

By setting the derivative of Eq. (18) respect to $\mathbf{F}$ to 0, we have

$$-\mathbf{W}+\mathbf{F}+\rho\left(\mathbf{F}\mathbf{F}^T\mathbf{F}-\mathbf{F}\right)=0, \tag{19}$$

then $\mathbf{F}$ can be updated by the following equation in each iteration:

$$\mathbf{F}_{ij} = \frac{\mathbf{W}_{ij}}{\left[\mathbf{F}+\rho\left(\mathbf{F}\mathbf{F}^T\mathbf{F}-\mathbf{F}\right)\right]_{ij}}. \tag{20}$$

4.3. Optimize Z. When fixing other variables, the optimization problem for $\mathbf{Z}$ is equal to the following equation:

$$\begin{aligned}\min_{\mathbf{Z}} &\left\|\mathbf{P}^T\mathbf{X}-\mathbf{F}\mathbf{Z}^T\right\|_F^2 + \gamma Tr\left(\mathbf{Z}^T\mathbf{L}\mathbf{Z}\right)\\ &\text{s.t.}\mathbf{Z}^T\mathbf{Z}=\mathbf{I},\ \mathbf{Z}\geq 0.\end{aligned} \tag{21}$$

We add a penalty term for the constraint $\mathbf{Z}^T\mathbf{Z}=\mathbf{I}$ and a Lagrange multiplier for the constraint $\mathbf{Z}\geq 0$. Then, the Lagrange function for Eq. (21) can be written as follows:

$$\begin{aligned}\mathscr{L}(\mathbf{Z},\alpha) = \min_{\mathbf{Z}} &\left\|\mathbf{P}^T\mathbf{X}-\mathbf{F}\mathbf{Z}^T\right\|_F^2 + \gamma Tr\left(\mathbf{Z}^T\mathbf{L}\mathbf{Z}\right)\\ &+\frac{\kappa}{4}\left\|\mathbf{Z}^T\mathbf{Z}-\mathbf{I}\right\| + Tr\left(\alpha\mathbf{Z}^T\right).\end{aligned} \tag{22}$$

By setting the derivative of Eq. (22) with respect to $\mathbf{Z}$ to 0, we have

$$-2\mathbf{X}^T\mathbf{P}\mathbf{F}+2\mathbf{Z}\mathbf{F}^T\mathbf{F}+\kappa\left(\mathbf{Z}\mathbf{Z}^T\mathbf{Z}-\mathbf{Z}\right)+\alpha=0. \tag{23}$$

According to the Kuhn-Tucker conditions $\alpha_{ij}\mathbf{Z}_{ij}=0$, we have

$$\mathbf{Z}_{ij} = \frac{\left[2\mathbf{X}^T\mathbf{P}\mathbf{F}+\kappa\mathbf{Z}\right]_{ij}}{\left[2\mathbf{Z}\mathbf{F}^T\mathbf{F}+\kappa\mathbf{Z}\mathbf{Z}^T\mathbf{Z}\right]_{ij}}. \tag{24}$$

Table 1: Statistics of the microarray datasets used in our experiments.

Datasets	#instance	#gene number	#class
Colon	62	20000	2
Lung	203	12600	5
Tumors-11	174	12533	11
CLL_SUB_111	111	11340	3
Breast	95	4869	3
GCM	198	16063	14

After we solve the resultant optimization problem as described by Eq. (1), we can measure the importance of each gene dimension by calculating the l_2-norm of each row of $\mathbf{P}$. We summarize the optimization procedure of the G^{3}CS model in Algorithm 1.

The proposed algorithm converges well with increasing iteration times. In our experiments, when the objective function value change between two continuous iteration times is very small, we stop the optimization process and obtain good results.

5. Experimental Results

In this section, extensive experiments are conducted on several real microarray datasets to validate the efficacy of the proposed G^{3}CS. In order to demonstrate that the gene subset selected by G^{3}CS can obtain better classification results, we use three kinds of classification algorithms including Support Vector Machine (SVM), Random Forest (RF), and k -nearest neighbor (KNN) to test the selected gene subset obtained by different previous gene selection methods.

5.1. Microarray Datasets. Six publicly available microarray datasets are used in our experiments, which are colon cancer (colon) [71], B-cell chronic lymphocytic leukemia (CLL SUB 111), breast, lung, tumors-11, and global cancer map (GCM) (1CLL SUB 111 and lung can be downloaded from: http://featureselection.asu.edu/datasets.php; breast and GCM can be downloaded from: http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi; tumors-11 can be downloaded from: http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html.) and are used to test the performance of the proposed G^{3}CS and

Table 2: Averaged classification accuracy (ACC ± SD) of different methods by using different classifiers (%) (the best results are marked in bold font).

Methods	Classifiers	CLL_SUB_111	Breast	Lung	Tumors-11	SRBCT	GCM
F-test	*k*-NN	55.90 ± 6.73	56.54 ± 9.36	87.61 ± 2.51	56.37 ± 3.53	94.98 ± 1.52	52.09 ± 6.37
	RF	57.97 ± 6.34	58.09 ± 9.01	86.88 ± 1.72	55.07 ± 4.48	95.17 ± 2.37	50.27 ± 6.51
	SVM	57.07 ± 6.37	57.40 ± 8.17	85.38 ± 2.34	57.20 ± 4.14	95.37 ± 1.46	51.07 ± 6.81
RLR	*k*-NN	75.97 ± 6.24	61.63 ± 7.27	91.37 ± 2.52	82.37 ± 3.47	96.04 ± 1.61	63.96 ± 5.25
	RF	73.10 ± 5.31	61.87 ± 7.32	91.69 ± 2.07	82.77 ± 4.35	97.09 ± 1.72	60.73 ± 5.34
	SVM	74.63 ± 5.45	60.11 ± 7.29	93.34 ± 2.18	81.23 ± 4.07	97.08 ± 1.38	61.79 ± 5.29
WLMGS	*k*-NN	73.58 ± 5.37	59.37 ± 8.07	91.17 ± 2.47	79.18 ± 4.27	97.08 ± 2.92	59.79 ± 4.56
	RF	74.76 ± 6.37	61.13 ± 7.51	91.68 ± 2.17	82.53 ± 4.41	96.95 ± 1.60	59.75 ± 4.67
	SVM	74.99 ± 6.74	59.33 ± 7.24	92.26 ± 2.37	80.88 ± 4.38	97.24 ± 1.53	61.48 ± 4.35
LNNFW	*k*-NN	75.34 ± 6.73	60.17 ± 7.42	89.39 ± 2.74	81.00 ± 4.94	95.76 ± 1.19	61.44 ± 5.19
	RF	73.86 ± 5.42	59.82 ± 7.41	92.43 ± 2.22	81.75 ± 4.37	96.51 ± 2.53	61.38 ± 5.61
	SVM	74.69 ± 6.14	60.37 ± 7.51	91.74 ± 2.84	81.91 ± 4.07	97.62 ± 2.80	62.57 ± 5.33
GRSL-GS	*k*-NN	76.19 ± 5.72	63.94 ± 7.70	93.47 ± 2.72	82.14 ± 4.84	97.88 ± 1.34	64.14 ± 4.63
	RF	76.37 ± 5.43	62.94 ± 7.30	94.02 ± 2.67	82.12 ± 4.62	97.46 ± 1.20	62.70 ± 5.24
	SVM	75.76 ± 5.34	62.45 ± 7.74	93.09 ± 2.33	82.96 ± 3.77	97.66 ± 1.74	63.74 ± 4.31
AHEDL	*k*-NN	76.97 ± 5.44	65.34 ± 7.64	93.48 ± 2.13	84.74 ± 4.96	98.37 ± 1.15	65.34 ± 4.34
	RF	76.88 ± 5.19	64.18 ± 0.74	95.12 ± 2.64	82.79 ± 4.33	98.13 ± 1.15	64.24 ± 5.34
	SVM	76.37 ± 5.04	65.87 ± 7.32	94.15 ± 2.31	83.07 ± 3.46	98.61 ± 1.48	65.57 ± 4.64
G3CS	*k*-NN	**78.37 ± 5.14**	66.75 ± 7.34	94.78 ± 2.45	**85.69 ± 4.32**	98.87 ± 1.25	66.88 ± 4.64
	RF	77.19 ± 5.69	65.77 ± 0.31	94.39 ± 2.85	84.89 ± 4.36	98.42 ± 1.25	66.04 ± 5.37
	SVM	77.67 ± 5.81	**66.96 ± 7.03**	**95.65 ± 2.18**	84.88 ± 3.59	**98.97 ± 1.38**	**67.11 ± 4.35**

other gene selection methods used for comparison. These datasets are collected for diagnosis of different kinds of cancers such as colon cancer, lung cancer, Ewing's family of tumors, non-Hodgkin lymphoma, and rhabdomyosarcoma and prostate cancer. For an instance, CLL SUB 111 contains high-density oligonucleotide arrays which can be used to identify molecular correlates of genetically and clinically distinct subgroups of B-cell chronic lymphocytic leukemia (B-CLL). Lung is a dataset used to determine whether global biological differences underlie common pathological features of prostate cancer and to identify genes that might anticipate the clinical behaviour of this disease.

It should be noted that the above six datasets are typical with high-dimensional genes. In each dataset, the number of genes is much larger than the number of samples, which brings challenge for many practical tasks. In Table 1, we give a brief description about these datasets.

5.2. Experimental Settings. In the proposed G^{3}CS, we have three parameters that need to be adjusted, i.e., α, β, and γ. In our experiments, we varied their values by a "grid-search" in the range $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. In addition, the optimal number of selected genes is also unknown, we set different numbers of selected genes for different datasets, and the final best results obtained from the optimal parameter setting were reported. In our experiments, the number of selected genes was tuned from {10,20,30,40,50} for each dataset. For each gene subset, the three abovementioned different basic classification methods were used to classify the microarray data for testing the discrimination of selected genes. In order to validate the efficacy of the proposed G^{3}CS, we compare it with other six state of-the-art gene selection methods including:

(i) *F*-test [72], which is a traditional filter-based gene selection method, it uses the statistical hypothesis testing

(ii) RLR [73], which is based on linear discriminant analysis criterion. The class centroid is estimated to define both the between-class separability and the within-class compactness

(iii) WLMGS (Weight Local Modularity based Gene Selection) [74], which uses the weight local modularity of a weighted sample graph to evaluate the discriminative power of gene subset

(iv) LNNFW [75], which uses the *k*-nearest neighbors rule to minimize the within-class distances and maximize the between-class distances

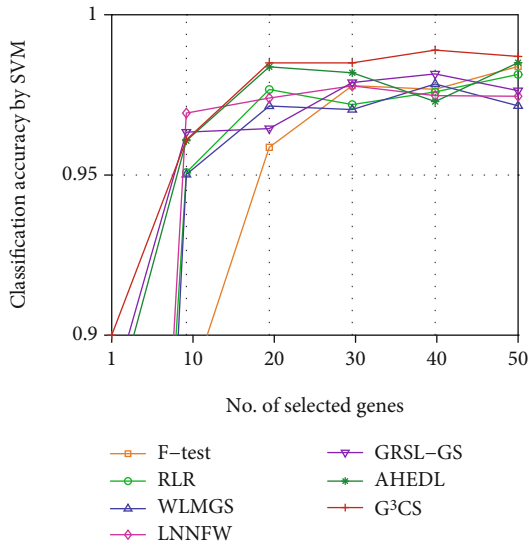

Figure 2: The classification accuracy of different methods with different selected number of genes on colon dataset.

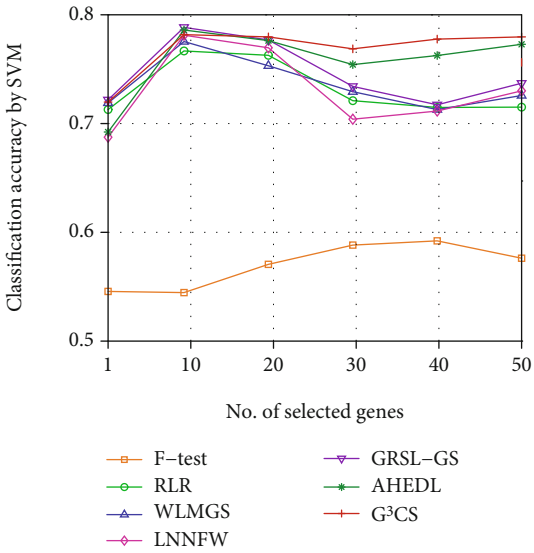

Figure 3: The classification accuracy of different methods with different selected number of genes on CLL SUB 111 dataset.

(v) GRSL-GS [20], which is based on subspace learning and manifold regularization

(vi) AHEDL [22], which is based on dictionary learning theory with adaptive hypergraph learning and regularization

As to WLMGS and GRSL-GS, we set the number of nearest neighbor for constructing the sample graph to 5. The kernel width σ used in the Gaussian kernel function and other regularization parameters in GRSL-GS and RLR are tuned with 5-fold cross validation (CV). For other parameters in other methods, we use the recommended settings in the corresponding references. We run all the implementation programs on a desktop computer with Intel Core i5-4200M 2.5 GHz CPU and 8 GB RAM.

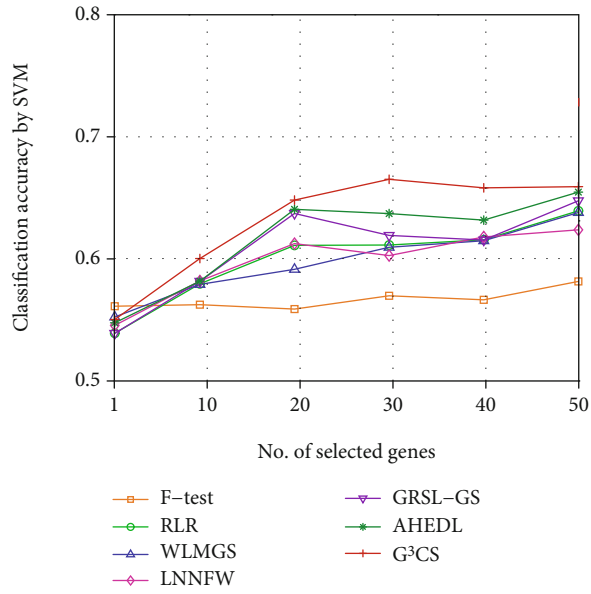

Figure 4: The classification accuracy of different methods with different selected number of genes on breast dataset.

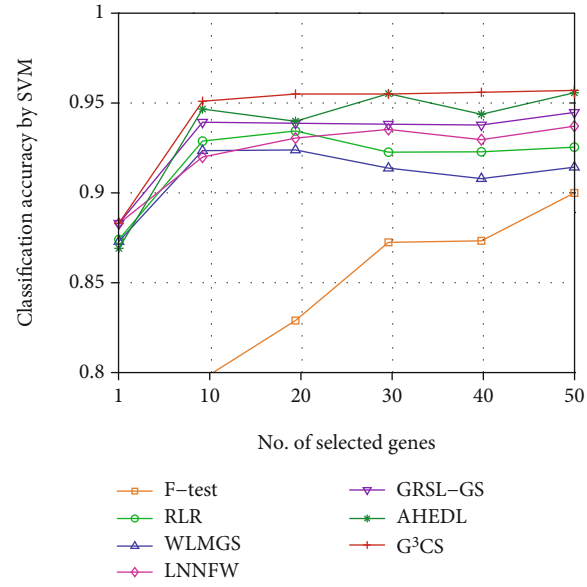

Figure 5: The classification accuracy of different methods with different selected number of genes on lung dataset.

5.3. Experimental Comparison of Different Methods. In order to verify the superiority of the proposed G^{3}CS, we compare it with the other six state-of-the-art gene selection methods on different datasets. For each dataset, we can obtain 5 different gene subsets with the numbers of selected genes which vary from 10 to 50. As to each gene subset, three classifiers and 5-fold CV are used for classification performance evaluation, and we report the average accuracy of 5 times of CV in Table 2. We mark the best results in bold font for clear

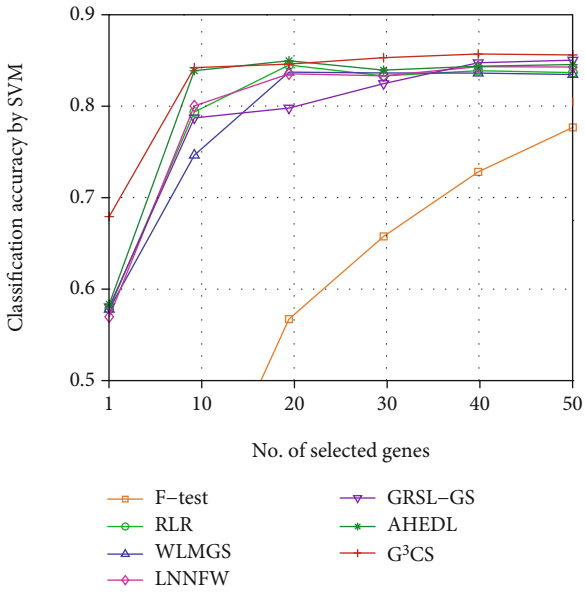

Figure 6: The classification accuracy of different methods with different selected number of genes on tumors-11 dataset.

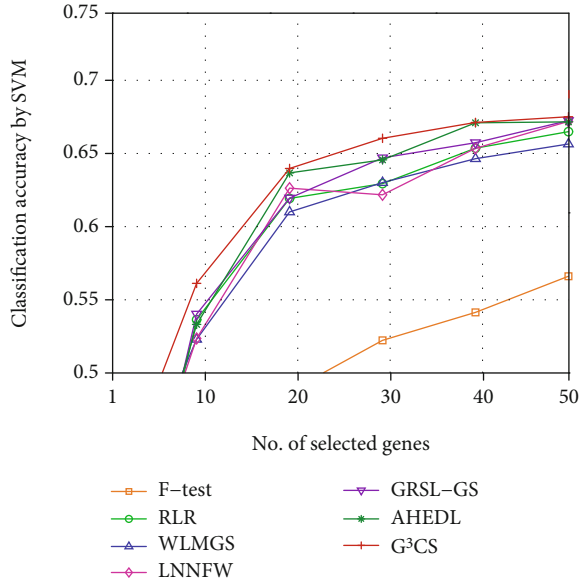

Figure 7: The classification accuracy of different methods with different selected number of genes on GCM dataset.

comparison. As can be seen from the results, the proposed G^{3}CS can consistently outperform other methods in terms of averaged classification accuracy, which demonstrates that G^{3}CS can effectively select more discriminative genes for original high-dimensional microarray data for classification task.

5.4. Classification Accuracy with Different Numbers of Selected Genes. Since the optimal number of selected genes for each dataset is hard to determine, we investigate the classification performance of different methods on different datasets with different numbers of selected genes. We plot the classification accuracy curves of different methods on different datasets with varied numbers of selected genes in Figures 2–7. For each method and each dataset, we plot the average classification accuracy value of the 5 times CV obtained by the SVM classifier. As can be seen from Figures 2–7, the proposed G^{3}CS performs steadily better than other methods when the number of selected genes changes. With a small number of selected genes, our method can select more discriminative genes than other methods, which validates that the selected gene subset obtained by G^{3}CS can better serve classification of microarray data.

6. Discussion and Conclusions

In this work, we present a novel gene selection method by taking the gene correlation into consideration, named gene correlation guided gene selection (G^{3}CS). In detail, we capture the correlation of different gene dimension pairs by calculating the covariance matrix from the perspective of gene dimension and embed it into the proposed model to regularize the gene selection coefficient learning. In such a manner, redundant genes can be effectively excluded to reduce the redundancy of the selected genes. In addition, a matrix factorization term was utilized to exploit the cluster structure of original microarray data to assist the learning process. We design an iterative updating algorithm to solve the resultant optimization problem. Experimental results on six publicly available microarray datasets are conducted to validate the efficacy of our proposed method. With varied selected gene dimensions, the proposed method can consistently outperform other compared ones in terms of classification accuracy.

Data Availability

The datasets used in this work are publicly available at: http://featureselection.asu.edu/datasets.php, http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We would like to thank Dr. Zheng for providing their Matlab code for generating the comparison results of this paper.

References

[1] V. T. V. Lj, H. Dai, V. D. V. Mj et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[2] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1/3, pp. 389–422, 2002.

[3] T. Nguyen and S. Nahavandi, "Modified ahp for gene selection and cancer classification using type-2 fuzzy logic," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 2, pp. 273–287, 2016.

[4] L. de Cecco, M. Giannoccaro, E. Marchesi et al., "Integrative mirna-gene expression analysis enables refinement of associated biology and prediction of response to cetuximab in head and neck squamous cell cancer," *Genes*, vol. 8, no. 1, p. 35, 2017.

[5] L. Naranjo, C. J. Pérez, J. Martín, and Y. Campos-Roca, "A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications," *Computer Methods and Programs in Biomedicine*, vol. 142, pp. 147–156, 2017.

[6] X. Huang, Y. Gao, B. Jiang, Z. Zhou, and A. Zhan, "Reference gene selection for quantitative gene expression studies during biological invasions: a test on multiple genes and tissues in a model ascidian *Ciona savignyi*," *Gene*, vol. 576, no. 1, pp. 79–87, 2016.

[7] A. C. Anauate, M. F. Leal, F. Wisnieski et al., "Identification of suitable reference genes for miRNA expression normalization in gastric cancer," *Gene*, vol. 621, pp. 59–68, 2017.

[8] S. Zhang, J. Wang, T. Ghoshal et al., "lncRNA gene signatures for prediction of breast cancer intrinsic subtypes and prognosis," *Genes*, vol. 9, no. 2, p. 65, 2018.

[9] H. H. Huang and Y. Liang, "Hybrid $L_{1/2+2}$ method for gene selection in the Cox proportional hazards model," *Computer Methods and Programs in Biomedicine*, vol. 164, pp. 65–73, 2018.

[10] S. Das, A. Rai, D. C. Mishra, and S. N. Rai, "Statistical approach for selection of biologically informative genes," *Gene*, vol. 655, pp. 71–83, 2018.

[11] J. Li, Y. Wang, T. Jiang, H. Xiao, and X. Song, "Grouped gene selection and multi-classification of acute leukemia via new regularized multinomial regression," *Gene*, vol. 667, pp. 18–24, 2018.

[12] A. Dabba, A. Tari, and S. Meftali, "Hybridization of moth flame optimization algorithm and quantum computing for gene selection in microarray data," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 2, pp. 2731–2750, 2021.

[13] D. W. Scott, *The Curse of Dimensionality and Dimension Reduction*, John Wiley & Sons, Inc., 2008.

[14] I. S. Oh, J. S. Lee, and B. R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1424–1437, 2004.

[15] K. Buza, "Classification of gene expression data: a hubness-aware semi-supervised approach," *Computer Methods and Programs in Biomedicine*, vol. 127, no. C, pp. 105–113, 2016.

[16] H. Song, X. Zhang, C. Shi, S. Wang, A. Wu, and C. Wei, "Selection and verification of candidate reference genes for mature microrna expression by quantitative rt-pcr in the tea plant (camellia sinensis)," *Genes*, vol. 7, no. 6, p. 25, 2016.

[17] J. Ramos, J. A. Castellanos-Garzón, A. González-Briones, J. F. de Paz, and J. M. Corchado, "An agent-based clustering approach for gene selection in gene expression microarray," *Interdisciplinary Sciences Computational Life Sciences*, vol. 9, no. 1, pp. 1–13, 2017.

[18] Y. Miao, H. Jiang, H. Liu, and Y. D. Yao, "An Alzheimers disease related genes identification method based on multiple classifier integration," *Computer Methods and Programs in Biomedicine*, vol. 150, pp. 107–115, 2017.

[19] W. Z. Wang, B. P. Yang, C. L. Feng et al., "Efficient sugarcane transformation via bar gene selection," *Tropical Plant Biology*, vol. 10, no. 2, pp. 75–85, 2017.

[20] C. Tang, L. Cao, X. Zheng, and M. Wang, "Gene selection for microarray data classification via subspace learning and manifold regularization," *Medical & Biological Engineering & Computing*, vol. 56, no. 7, pp. 1271–1284, 2018.

[21] Z. Y. Algamal, R. Alhamzawi, and H. T. Mohammad Ali, "Gene selection for microarray gene expression classification using Bayesian lasso quantile regression," *Computers in Biology and Medicine*, vol. 97, pp. 145–152, 2018.

[22] X. Zheng, W. Zhu, C. Tang, and M. Wang, "Gene selection for microarray data classification via adaptive hypergraph embedded dictionary learning," *Gene*, vol. 706, pp. 188–200, 2019.

[23] J. Liu, R. Su, J. Zhang, and L. Wei, "Classification and gene selection of triple-negative breast cancer subtype embedding gene connectivity matrix in deep neural network," *Briefings in Bioinformatics*, 2021.

[24] E. P. Kirk, R. Ong, K. Boggs et al., "Gene selection for the Australian Reproductive Genetic Carrier Screening Project ("Mackenzie's Mission")," *European Journal of Human Genetics*, vol. 29, no. 1, pp. 79–87, 2021.

[25] C. Tang, X. Zheng, X. Liu et al., "Crossview locality preserved diversity and consensus learning for multi-view unsupervised feature selection," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[26] P. Mitra, C. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 301–312, 2002.

[27] M. al-Rajab, J. Lu, and Q. Xu, "Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis," *Computer Methods and Programs in Biomedicine*, vol. 146, pp. 11–24, 2017.

[28] H. Shi, Y. Luo, C. Xu, and Y. Wen, "Manifold regularized transfer distance metric learning," in *British Machine Vision Conference*, pp. 158.1–158.11, Swansea, UK, 2015.

[29] X. Shen, F. Shen, L. Liu, Y. Yuan, W. Liu, and Q. Sun, "Multi-view discrete hashing for scalable multimedia search," *ACM Transactions on Intelligent Systems and Technology*, vol. 9, no. 5, pp. 53:1–53:21, 2018.

[30] X. Shen, F. Shen, Q. S. Sun, Y. Yang, Y. H. Yuan, and H. T. Shen, "Semi-paired discrete hashing: learning latent hash codes for semi-paired cross-view retrieval," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4275–4288, 2017.

[31] S. Li, C. Tang, X. Liu, Y. Liu, and J. Chen, "Dual graph regularized compact feature representation for unsupervised feature selection," *Neurocomputing*, vol. 342, no. 331, pp. 77–96, 2019.

[32] C. Tang, X. Zhu, J. Chen, P. Wang, X. Liu, and J. Tian, "Robust graph regularized unsupervised feature selection," *Expert Systems with Applications*, vol. 96, pp. 64–76, 2018.

[33] C. Tang, X. Liu, M. Li et al., "Robust unsupervised feature selection via dual self-representation and manifold regularization," *Knowledge-Based Systems*, vol. 145, pp. 109–120, 2018.

[34] C. Tang, J. Chen, X. Liu et al., "Consensus learning guided multi-view unsupervised feature selection," *Knowledge-Based Systems*, vol. 160, pp. 49–60, 2018.

[35] C. Tang, X. Zhu, X. Liu et al., "Learning a joint affinity graph for multiview subspace clustering," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1724–1736, 2018.

[36] C. Tang, X. Zhu, X. Liu, and L. Wang, "Cross-view local structure preserved diversity and consensus learning for multi-view unsupervised feature selection," in *AAAI Conference on Artificial Intelligence*, pp. 5101–5108, Hawaii, USA, 2019.

[37] C. Tang, X. Liu, X. Zhu et al., "Feature selective projection with low-rank embedding and dual Laplacian regularization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 9, pp. 1747–1760, 2019.

[38] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments," *Statistica Sinica*, vol. 12, no. 1, pp. 111–139, 2000.

[39] A. D. Long, H. J. Mangalam, B. Y. Chan, L. Tolleri, G. W. Hatfield, and P. Baldi, "Improved Statistical Inference from DNA Microarray Data Using Analysis of Variance and A Bayesian Statistical Framework," *Journal of Biological Chemistry*, vol. 276, no. 23, pp. 19937–19944, 2001.

[40] J. G. Thomas, J. M. Olson, S. J. Tapscott, and L. P. Zhao, "An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles," *Genome Research*, vol. 11, no. 7, pp. 1227–1236, 2001.

[41] T. Golub, D. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[42] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," *Advances in Neural Information Processing Systems*, vol. 18, pp. 507–514, 2005.

[43] R. Cai, Z. Hao, X. Yang, and W. Wen, "An efficient gene selection algorithm based on mutual information," *Neurocomputing*, vol. 72, no. 4-6, pp. 991–999, 2009.

[44] L. Y. Chuang, C. H. Yang, J. C. Li, and C. H. Yang, "A hybrid bpso-cga approach for gene selection and classification of microarray data," *Journal of Computational Biology*, vol. 19, no. 1, pp. 68–82, 2012.

[45] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine Learning*, vol. 53, no. 1/2, pp. 23–69, 2003.

[46] Hanchuan Peng, Fuhui Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[47] Z. Yi, C. Ding, and L. Tao, "Gene selection algorithm by combining relieff and mrmr," *BMC Genomics*, vol. 9, no. S2, p. S27, 2008.

[48] H. Kim, S.-M. Choi, and S. Park, "Gseh: a novel approach to select prostate cancer-associated genes using gene expression heterogeneity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 1, pp. 129–146, 2018.

[49] K. B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple svm-rfe for gene selection in cancer classification with expression data," *IEEE Transactions on Nanobioscience*, vol. 4, no. 3, pp. 228–234, 2005.

[50] X. Zhou and D. P. Tuck, "Msvm-rfe: extensions of svm-rfe for multiclass gene selection on dna microarray data," *Bioinformatics*, vol. 23, no. 9, pp. 1106–1114, 2007.

[51] Y. Liang, F. Zhang, J. Wang, T. Joshi, Y. Wang, and D. Xu, "Prediction of drought-resistant genes in arabidopsis thaliana using svm-rfe," *PLoS One*, vol. 6, no. 7, article e21750, 2011.

[52] E. Tapia, P. Bulacio, and L. Angelone, "Sparse and stable gene selection with consensus svm-rfe," *Pattern Recognition Letters*, vol. 33, no. 2, pp. 164–172, 2012.

[53] X. Han, X. Chang, L. Quan et al., "Feature subset selection by gravitational search algorithm optimization," *Information Sciences*, vol. 281, pp. 128–146, 2014.

[54] D. Ghosh and A. M. Chinnaiyan, "Classification and selection of biomarkers in genomic data using lasso," *Journal of Biomedicine and Biotechnology*, vol. 2005, no. 2, 154 pages, 2005.

[55] G. Wu, R. Mallipeddi, and P. N. Suganthan, "Ensemble strategies for population-based optimization algorithms - a survey," *Swarm and Evolutionary Computation*, vol. 44, pp. 695–711, 2019.

[56] A. K. Shukla, "Identification of cancerous gene groups from microarray data by employing adaptive genetic and support vector machine technique," *Computational Intelligence*, vol. 36, no. 1, pp. 102–131, 2020.

[57] S. Das, A. Abraham, U. K. Chakraborty, and A. Konar, "Differential evolution using a neighborhood-based mutation operator," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 3, pp. 526–553, 2009.

[58] S. Dwivedi, M. Vardhan, and S. Tripathi, "Incorporating evolutionary computation for securing wireless network against cyberthreats," *The Journal of Supercomputing*, vol. 76, pp. 8691–8728, 2020.

[59] P. Zhu, W. Zuo, L. Zhang, Q. Hu, and S. C. K. Shiu, "Unsupervised feature selection by regularized self-representation," *Pattern Recognition*, vol. 48, no. 2, pp. 438–446, 2015.

[60] R. Shang, Z. Zhang, L. Jiao, C. Liu, and Y. Li, "Self-representation based dual-graph regularized feature selection clustering," *Neurocomputing*, vol. 171, no. 1, pp. 1242–1253, 2016.

[61] P. Zhu, W. Zhu, W. Wang, W. Zuo, and Q. Hu, "Non-convex regularized self-representation for unsupervised feature selection," *Image and Vision Computing*, vol. 60, no. 1, pp. 22–29, 2017.

[62] Y. Liu, K. Liu, C. Zhang, J. Wang, and X. Wang, "Unsupervised feature selection via diversity-induced self-representation," *Neurocomputing*, vol. 219, pp. 350–363, 2017.

[63] Y. X. Wang, J. X. Liu, Y. L. Gao, C. H. Zheng, and J. L. Shang, "Differentially expressed genes selection via Laplacian regularized low-rank representation method," *Computational Biology and Chemistry*, vol. 65, no. 1, pp. 185–192, 2016.

[64] R. Zheng, M. Li, Z. Liang, F.-X. Wu, Y. Pan, and J. Wang, "Sinnlrr: a robust subspace clustering method for cell type detection by non-negative and low-rank representation," *Bioinformatics*, vol. 35, no. 19, pp. 3642–3650, 2019.

[65] C.-H. Zheng, T.-Y. Ng, L. Zhang, C.-K. Shiu, and H.-Q. Wang, "Tumor classification based on non-negative matrix factorization using gene expression data," *IEEE Transactions on Nanobioscience*, vol. 10, no. 2, pp. 86–93, 2011.

[66] D. Wang, J. X. Liu, Y. L. Gao, J. Yu, C. H. Zheng, and Y. Xu, "An NMF-L2,1-norm constraint method for characteristic gene selection," *PLoS One*, vol. 11, no. 7, article e0158494, 2016.

[67] S. Du, Y. Ma, S. Li, and Y. Ma, "Robust unsupervised feature selection via matrix factorization," *Neurocomputing*, vol. 241, pp. 115–127, 2017.

[68] X. Guo, X. Jiang, J. Xu, X. Quan, M. Wu, and H. Zhang, "Ensemble consensus-guided unsupervised feature selection to identify Huntington's disease-associated genes," *Genes*, vol. 9, no. 7, p. 350, 2018.

[69] D. Han and J. Kim, "Unsupervised simultaneous orthogonal basis clustering feature selection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5016–5023, Boston, USA, 2015.

[70] M. Qi, T. Wang, F. Liu, B. Zhang, J. Wang, and Y. Yi, "Unsupervised feature selection by regularized matrix factorization," *Neurocomputing*, vol. 273, pp. 593–610, 2018.

[71] U. Alon, N. Barkai, D. A. Notterman et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[72] K. A. Lê Cao, A. Bonnet, and S. Gadat, "Multiclass classification and gene selection with a stochastic algorithm," *Computational Statistics and Data Analysis*, vol. 53, no. 10, pp. 3601–3615, 2009.

[73] S. Guo, D. Guo, L. Chen, and Q. Jiang, "A centroid-based gene selection method for microarray data classification," *Journal of Theoretical Biology*, vol. 400, pp. 32–41, 2016.

[74] G. Zhao and Y. Wu, "Feature subset selection for cancer classification using weight local modularity," *Scientific Reports*, vol. 6, no. 1, p. 34759, 2016.

[75] S. An, J. Wang, and J. Wei, "Local-nearest-neighbors-based feature weighting for gene selection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 5, pp. 1538–1548, 2018.