PLS-Based Gene Selection and Identification of Tumor-Specific Genes

Guoli Ji, Zijiang Yang, and Wenjie You

***Abstract*—In view of the characteristics of high-dimensional small sample, strong relevance, and high noise of the identification of tumor-specific genes on microarray, a novel partial least squares (PLS) based gene-selection method, which synthesizes genetic relatedness and is suitable for multicategory classification, is presented. Using the explanation difference of independent variables on dependent variable (class), we define three indicators for global gene selection, which takes into accounts the combined effects of all the genes and the correlation among the genes. Integrated with the linear kernel support vector classifier (SVC), the proposed method is tested by MIT acute myeloid leukemia/acute lymphoblastic leukemia (AML/ALL) and small round blue cell tumors (SRBCT) data sets. A subset of specific genes with small numbers and high identification are obtained. The results indicate that our proposed PLS-based method for tumor-specific genes selection is highly efficient. Compared to the literature, the selected specific genes from both two-category dataset AML/ALL and multicategory dataset SRBCT are credible. Further investigation shows that the proposed gene-selection method is robust. Overall, the proposed method can effectively solve feature-selection problem on high-dimensional small sample. At the same time, it has good performance for multicategory classification as well.**

***Index Terms*—Gene selection, high-dimensional small samples, partial least squares (PLS), tumor-specific gene.**

I. Introduction

TUMOR classification and diagnosis of gene-expression profiles is a popular topic in the field of bioinformatics. In modern biomedicine, the occurrence and development of tumors is reflected by differences in tumor-related gene expression to a certain extent. A method for cancer treatment and prevention can be found through the identification of tumor-specific genes and their functions. The study by Golub *et al.* [8] showed that differences in gene expression of tumor subtypes in clinical diagnosis can be specific to the detection of gene expression and treatment. How to isolate tumor-specific genes from a number of tumor-related genes is the problem of tumor-specific gene selection.

The challenge of gene-chip analysis and studies is to create an effective mathematical model to deal with such a small sample with large number of genes [8], [11]. The essence of the small sample with large number of genes lies in the information redundancy and high noise. Effective modeling of small samples requires one to retain a maximum amount of useful information, while simultaneously removing redundancy and noise. Many data mining algorithms lack efficiency or even fail in such cases. The common method used to address the problem of high-dimension small sample is to compress its feature dimension. There are two dimensionality reduction methods: feature selection (dimensionality selection) and feature extraction (dimensionality reduction). Therefore, one research direction is the building of effective feature selection for high-dimension data mining. Effective gene selection is more important than the classifier in the analysis of gene microarray data.

In recent years, support vector machine (SVM), which was developed on the basis of statistical learning theory, is employed to accomplish feature selection on high-dimension small data set, and has achieved good results. However, there is an obvious limitation on the choice of kernel function and kernel parameters. This paper proposes the partial least squares (PLS), which can effectively eliminate the effects of collinearity between variables, and has equally efficient and strong predictive power in handling the small samples with large number of genes. Using explanation difference of independent variables on dependent variable (class), we define three indicators [(independent-variable explanation gain (IEG), dependent-variable explanation gain (DEG), and variant importance in projection (VIP)]. In addition, a new filter-based method of global gene selection, where each specific gene is extracted based on all sample genes in the input domain is proposed by measuring classification information of each variable, which synthesizes genetic relatedness and is suitable for multicategory classification. Furthermore, our novel PLS-based global gene selection takes into account the correlation of all genes. It can detect those genes with a relatively small main effect, but with a strong interaction effect. Our proposed gene-selection algorithm opens a new way for tumor-specific genes selection. The MIT acute myeloid leukemia/acute lymphoblastic leukemia (AML/ALL) and small round blue cell tumors (SRBCT) data sets are used to validate our proposed method, and a subset of specific genes with small numbers and high identification are gained. The results indicate that our PLS-based tumor-specific

Manuscript received March 3, 2010; revised July 7, 2010; accepted September 7, 2010. Date of publication October 21, 2010; date of current version October 19, 2011. This work was supported in part by the National Natural Science Foundation of China under Project 60774033, in part by the Special Research Fund for the Doctoral Program of Higher Education under Project 20070384003 and Project 20090121110022, in part by Xiamen University under National 211 Project 0630-E62000, and in part by the Natural Sciences and Engineering Research Council of Canada. This paper was recommended by Associate Editor R. Alhajj.

G. Ji (corresponding author) is with the Department of Automation, Xiamen University, Xiamen, Fujian 361005, China (e-mail: glji@xmu.edu.cn).

Z. Yang is with the School of Information Technology, York University, Toronto M3J 1P3, Canada (e-mail: zyang@mathstat.yorku.ca).

W. You is with the Department of Mathematics and Computer Science, Fujian Normal University, Fujian 350300, China. He is also with the Department of Automation, Xiamen University, Xiamen, Fujian 361005, China (e-mail: ywj_huang@163.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCC.2010.2078503

genes selection is highly efficient. Compared to other studies in literature, the selected specific genes from both two-category dataset AML/ALL and multicategory dataset SRBCT are credible.

The rest of the paper is organized as follows. A literature review is provided in Section II that discusses the background of traditional methods of gene selection. Section III presents the proposed PLS-based indicators for new filter-based global gene-selection method through PLS model. The approach is applied to two-category AML/ALL and multicategory SRBCT, and the results are compared with the related literature, and further performance evaluation is presented in Section IV. Finally, conclusion and future work are given in Section V.

II. Literature Review

Feature selection utilizes standard statistical indicators or separation criterion to select some features with the largest contribution to the classification. Once its dimensionality has been reduced, the main features of the original data are preserved, i.e., from the original data table, a number of notable features relevant to the tasks are selected and a new low-dimension data table is created. The feature selection results in no rotation of the data table, and the results are easy to explain. Tumor-specific gene selection is a feature-selection problem of pattern recognition. Its goal is to obtain a subset of specific genes with small number and high identification. Usually two types of methods can be used in feature selection: filter methods and wrapper methods. Filter method is a single gene-scoring method based on certain criteria, which usually selects high-score genes for further analysis. The most popular methods include t-statistic, signal-to-noise ratio (SNR) [2], [8], [16], nonparametric rank sum statistic [3], Markov Blanket filter [30], mutual information [9], [24], and Relief algorithm and its improvement [12], [14]. Wrapper method is a feature-selection method associated with classifier. The output of the classifier is treated as a feature-selection criterion. Wrapper methods include SNR, recursive feature elimination (RFE) Relief [26] method, and SVMs [11], [21], [22] and RFE method, etc.

The related literature on feature selection based on DNA microarray includes Golub's t-statistic and SNR and its revised version [18], Khan *et al.*'s [13] artificial neural network, Tibshirani *et al.*'s [27] nearest shrunken centroid, Deng *et al.*'s [3] nonparametric rank sum test, Feng and Shi [6] improved Fisher model, and Ruan *et al.*'s improved RFE-Relief algorithm [26]. Wrapper methods include Li and Ruan [18] improved SNR method, RFE-Relief method, integrate IIC and redundancy with sensitivity analysis, and Guyon *et al.*'s [11] SVMs and RFE method, Li and Yang's [19] RR-RFE method, M. Banerjee *et al.*'s [1] evolutionary rough method, Maldonado and Weber [21], who proposed Hold-Out SVM (HO-SVM) and so on.

PLS approach has been applied in the data analysis of gene-expression profiles. Nguyen and Rocke [23] and Dai *et al.* [2] applied PLS as a dimension-reduction method named as PLSDR instead of a regression model. Li *et al.* proposed t-statistic score-based PLSDR for redundant feature selection and irrelevant feature selection in two-category classification problem [16], [31], [32]. Gutkin *et al.* [10] presented a PLS-based feature-selection method dealing with two-category classification.

In order to speed up the gene-selection process of the tumor-gene expression, ranking is usually done to microarray data of high dimension and small sample based on the value of test statistics on single gene (variable) containing the information of classification such as Golub *et al.*'s signal-to-noise (S2N) and *t-test* and its p value in statistics [2], [8], [16], [18], [27]. There may be a risk in this method, which ignores the relationship and nonlinearity between genes. A more accurate approach needs to consider the joint distribution between genes. It must take into account all of the genes and allow detection of those with smaller main effects, but with the strong interaction effects of genes. Because these genes, which have a smaller main effect, may be coregulating genes on the biochemical pathways of organisms, the analysis of these genes may provide more comprehensive understanding of the characteristics of tumor-specific gene expression.

The relationship between the features is not considered in the single-gene scoring method. The Relief algorithm, ridge regression, SVM-based gene selection and its improvement, and the RFE algorithm [11], [19], [21], [26] take into account the correlation between genes to a certain extent. The combination of wrapper feature selection [11], [21] on SVMs, and supervised classification on SVMs was one of the better-performing methods. However, since the optimal kernel function and its corresponding kernel parameters in each data set are not fixed in SVMs, the kernel function and its parameters need to be reestablished using the new data sets. As other machine-learning methods (such as neural networks), SVMs is also a black-box technology. Thus, the transformation from low-dimension to high-dimension data makes it difficult to explain the classification process. In addition, although, some gene-selection algorithms can find a small subset of most predictive genes, they are very time consuming.

In this paper, PLS-based global gene-selection method is proposed. It considers the correlation between genes, while its computational complexity is less than the traditional S2N and t-test methods. In high-throughput gene microarray, the new filter method can be used to obtain a selection of information genes from an entire range in order to reduce the search space. Then subsets of information genes that meet the conditions are selected by wrapper method. Global feature selection takes into account the correlation of all genes. It can detect those genes with a relatively small main effect, but with a strong interaction effect. Furthermore, the computational complexity of simple partial least squares (SIMPLS) is $O(np)$ [2], [10], [16], a linear function of p (gene number). Thus, it is very efficient, which is particularly obvious on the high-dimension small samples of gene microarray data. Finally, our feature-selection algorithm can accomplish both two-category classification and multicategory classification, while many existing methods (such as t-test, SNR, etc.) are only for two-category classification, and have limitations in multiple classification.

III. Research Methodology

A. Tumor Data

1) MIT AML/ALL Data: Leukemia is a malignant tumor of the hematopoietic system, manifested as a malignancy of one or more components of blood cells in a hematopoietic system, such as bone marrow or lymph nodes. Body tissues are immerged leading to inhibition of normal hematopoietic cells. Then, a variety of symptoms are exhibited. The leading incidence of tumors in children is acute leukemia. According to the form of leukemia cells and the staining performance, the diseases were divided into two types clinically: ALL and AML. Their main clinical manifestations were similar in spite of types of acute leukemia cells, and the early symptoms may not be apparent because the symptoms were similar to common childhood diseases. Therefore, accurate classification between ALL and AML is very useful in early diagnosis on acute leukemia, targeted therapy, and the improvement of survival and quality of life.

Using high-density oligonucleotide arrays, Golub *et al.* [8] and others detected 7129 gene-expression levels with a training data set containing 38 samples (27 ALL, 11 AML) and a test data containing 34 samples (20 ALL, 14 AML), which can be downloaded from http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi. Golub *et al*. filtered 50 genes, and structured a classification according to 38 training samples. They applied their results to 34 new test samples, and 29 samples were correctly identified. Guyon *et al*. [11] used SVMs and RFE for feature selection and classification. He selected 8 and 16 genes. The training samples and test samples were all correctly classified. Feng and Shi [6] used the fast fisher optimization model to select 16 genes. Training samples and test samples were all correctly identified. Li *et al*. [17] adopted model-free gene-selection method by considering unbalanced samples. Thirty genes were selected by SVMs. Training samples and test samples were all correctly identified. Li and Ruan [18] used IIC, redundancy, and sensitivity. Sixteen genes were selected by SVMs. Training samples and test samples were all correctly identified. Li and Yang [19] used RR-RFE to identify three significant genes with 100% correct classification on test samples. Our algorithm selected eight and nine genes, only using a simple linear kernel support vector classifier (SVC) (default parameters) to identify 100% of training samples and test samples.

2) SRBCT Data: SRBCT usually occurs in children who rarely survive beyond the age of 30. SRBCT actually has four different types of tumors: Ewing's sarcoma (EWS), Burkett's lymphoma (BL), neuroblastoma (NB), and rhabdomyosarcoma (RMS). Under a microscope, these tumor cells are alike, which makes them extremely difficult to be identified quickly and accurately by traditional diagnostic methods. Khan *et al*. detected 2308 levels of gene expression of SRBCT [13]. It contains 2308 genes. All samples are divided into a training set with 63 samples and a test set with 20 samples, which can be downloaded from http://research.nhgri.nih.gov/microarray/Supplement/. Khan *et al*. [13] selected 96 genes using an artificial neural network. The training samples were classified correctly. However, the diagnostic accuracy is only 90%. Tibshirani *et al*. [27] selected 43 genes. The test samples were correctly classified. Li *et al*. [17] adopted model-free gene-selection method by considering unbalanced samples. Seventy-four genes were selected. Both the training samples and test samples were all correctly identified. Our algorithm selected 15 and 24 genes, where only using simple linear kernel SVC (default parameters) is enough to identify 100% of training samples and test samples.

B. Models and Algorithms

1) Partial Least Squares (PLS): Let X be the $n \times p$ matrix of n tissue samples and p genes, and Y be the $n \times q$ matrix of q classes. Note that X and Y used here are assumed to be centered to zero mean by each column. The goal of PLS is to find a pair of projection directions (weight vectors) w and c so that the projections (PLS components) $t = Xw$ and $u = Yc$ meets the following criteria.

1) Let t and u load variation information of X and Y as much as possible, i.e.,

$$\mathrm{Var}(t) \to \mathrm{max.and.Var}(u) \to \mathrm{max}.$$

2) The correlation coefficient is maximized for t and u, i.e.,

$$r(t, u) \to \mathrm{max}.$$

Integrating 1) and 2), PLS requests to maximize the covariance of t and u, i.e.,

$$\mathrm{Cov}(t, u) = \sqrt{\mathrm{Var}(t)\mathrm{Var}(u)}\, r(t, u) \to \mathrm{max}.$$

In addition

$$\begin{aligned}\mathrm{Cov}(t, u) &= E(t \cdot u) = E(w^T X Y^T c) = w^T E(XY^T) c\\ &= w^T S_{XY} c\end{aligned}$$

where S_{XY} is the cross-covariance matrix of $X$and Y. Therefore, the determination of PLS projection directions of w and c can be realized by maximizing the objective function

$$J(w, c) = w^T S_{XY} c = \frac{w^T S_{XY} c}{\sqrt{(w^T w)(c^T c)}}$$

where $w^T w = 1,\ c^T c = 1$.

Based on the aforementioned mathematical model, two projection directions w and c are determined by maximizing the objective function $J(w, c)$. Then, the data matrix X (predictor variables) and Y (response variable) are projected on the directions of w and c, respectively. Thus, the first pair of PLS composition $t_1 = Xw$ and $u_1 = Yc$ are obtained. Next, the regression equations with Y and t_1, the equations with Y and u_1, and the equations with X and t_1 are established. The algorithm terminates if the regression equations meet the accuracy requirements. Otherwise, the second latent variable t_2 was extracted from the residuals, which has been interpreted by t_1 of X, and u_2 was extracted from the residuals, which has been interpreted by t_1 of Y. Repeat this process until it reaches the required precision.

Therefore, the basic mathematical model of PLS can be formulated as follows:

$$\begin{cases} \max & \operatorname{cov}(Xw_i, Yc_i) \\ \text{s.t.} & w_i' w_i = 1, c_i' c_i = 1 \\ & w_i' \sum_X w_j = 0 \\ & c_i' \sum_Y c_j = 0 \end{cases}$$

where the linear combination $t_i = Xw_i$ is the ith latent variables, $\sum_X = X'X$, and $\sum_Y = Y'Y$.

The solution of the aforementioned optimization problem (w_i, c_i) can be found as follows [20], [29]:

$$w_i = \begin{cases} \Sigma_{XY}\Sigma_{YX} \text{ main eigenvector}, & i = 1 \\ (I - P_X)\Sigma_{XY}(I - P_Y)\Sigma_{YX} \text{ main eigenvector}, & i > 1 \end{cases}$$

$$c_i = \begin{cases} \Sigma_{YX} w_i, & i = 1 \\ (I - P_Y)\Sigma_{YX} w_i, & i > 1 \end{cases}$$

where

$$P_X = (\Sigma_X W)[(\Sigma_X W)^T (\Sigma_X W)]^{-1} (\Sigma_X W)^T$$

and

$$P_Y = (\Sigma_Y C)[(\Sigma_Y C)^T (\Sigma_Y C)]^{-1} (\Sigma_Y C)^T$$
$$W = (w_{ij})\, C = (c_{ij}).$$

Assuming $T = XW$ and $U = YC$, the PLS simultaneous matrix equations are formulated as follows:

$$\begin{cases} U = TB^T \\ X = TP^T + E \\ Y = UG^T + F = TQ^T + F \end{cases}$$

where T and U are score matrices, and P and Q are loading matrices for X and Y, respectively. Please refer to [25] and Appendix I for more detailed definitions of score matrices and loading matrices. Matrix B contains the PLS regression coefficients. E and F are residuals of X and Y, respectively.

In general, PLS regression models can be denoted by the form of matrix [25]

$$Y = XB + F$$

where $B = X^T U (T^T X X^T U)^{-1} T^T Y$ is the regression coefficient matrix and F is residuals matrix.

Basic PLS algorithms mainly include Wold's [28] nonlinear iterative partial least squares (NIPALSs) and de Jong's [4] SIMPLSs. SIMPLS is much faster than the NIPALS algorithm, especially when the dimension of the variable X increases. But, it gives slightly different results in the case of multivariate Y. Algorithm 1 gives de Jong's [4] classic SIMPLS algorithm. The computational complexity of SIMPLS is $O(np)$ [2], [10], [16], a linear polynomial of p (gene number). Thus, it is a highly efficient algorithm.

2) Proposed PLS-Based Global Gene-Selection Algorithm: First, we encode the dependent variable Y (class) in order to assure the category label independence, and implement the unified feature-selection algorithms, which accomplish two-category classification and multicategory classification simultaneously.

```
Algorithm 1   SIMPLS($X_{n\times p}$, $Y_{n\times m}$, nfac)
Input:   $X_{n\times p}$,   $Y_{n\times m}$,   nfac
Output:  T , U , P , Q , W , VarX , VarY

Initialization : Standardization of X, Y , $A(1) \leftarrow X'Y$ , $M(1) \leftarrow X'X$ , $E(1) \leftarrow I$
For $k \leftarrow 1$ to nfac
    $c(k) \leftarrow A'(k)A(k)$ dominant eigenvector
    $w(k) \leftarrow A(k)c(k)$
    $D(k) \leftarrow w(k)'Mw(k)$
    $w(k) \leftarrow w(k)/\mathrm{sqrt}(D(k))$
    $p(k) \leftarrow M(k)w(k)$
    $q(k) \leftarrow M(k)w(k)$
    $G(k) \leftarrow E(k)p(k)$
    $G(k) \leftarrow G(k)/\| G(k)\|$
    $E(k+1) \leftarrow E(k)\text{-}G(k)G(k)'$
    $M(k+1) \leftarrow M(k)\text{-}p(k)\,p(k)'$
    $A(k+1) \leftarrow E(k)\,A(k)$
Endfor
$T \leftarrow XW$ , $B \leftarrow WQ$
$VarX \leftarrow \mathrm{DIAG}(P'P)/(n\text{-}1)$ , $VarY \leftarrow \mathrm{DIAG}(Q'Q)/(n\text{-}1)$
```

Then, we define the three indicators used to achieve the PLS-based global gene-selection algorithm.

Essentially, PLS is a regression (fitting) method. Observations of independent variables predict the dependent variable. It is a typical quantitative prediction, which fits the continuous numerical data. But the dependent variable of classification is a qualitative problem. The independent variable selection was realized by the explanation of the dependent variable on PLS, and classification (two category, multicategory) analysis was fulfilled. The dependent variable Y is encoded [23] in order to achieve category label independence. The definition of category labels $Y = (y_{ij})_{n\times g}$ (n observed sample, g categories) is

$$y_{ij} = I(y_i = j) = \begin{cases} 1 & y_i = j \\ 0 & y_i \neq j \end{cases},\ i = 1, \ldots, n, j = 1, \ldots, g.$$

We define three indicators: IEG, DEG, and VIP. And the PLS-based gene-selection algorithm is presented in the following.

Definition 1 (Variable Variation Explanation): Given the component t_h and dependent variable y_k, we define the following:

$$Rd(y_k, t_h) = r^2(y_k, t_h)$$

as variable variation explanation between component t_h and dependent variable y_k, where $r(x, y)$ is the correlation coefficient of the two variables.

Definition 2 (Accumulation of Variation Explanation): Given the component $t_1, t_2, \ldots, t_m$ and the dependent variable Y, we define the following:

$$Rd(Y; t_1, \ldots, t_m) = \sum_{h=1}^{m} Rd(Y, t_h)$$

as accumulation of variation explanation between component $t_1, t_2, \ldots, t_m$ and the dependent variable Y.

The ith independent variable x_i ($i = 1, \ldots, p$) is left out and the remaining $p - 1$ variables are remodeled by PLS. We define $Rd_{(-i)}(X; t_1', \ldots, t_m')(Rd_{(-i)}(Y; t_1', \ldots, t_m'))$ as the accumulation of variation explanation between the new component $t_1', t_2', \ldots, t_m'$ and the dependent variable X (independent variable Y).

```
Algorithm 2  $PLSVEG(TrainX_{n\times p}, ClassY_{n\times g}, nfac)$
Input:   $TrainX_{n\times p} = (x_1, x_2, \ldots x_p)$ , $ClassY_{n\times g}$ , $nfac$
Output:  $ieg$ , $deg$

Call $SIMPLS(TrainX_{n\times p}, ClassY_{n\times g}, nfac)$ to get $Rd(X)$ , $Rd(Y)$ accumulation of
variation explanation of $X$, $Y$
For $i \leftarrow 1$ to $p$
    $TrainX_{(-i)} \leftarrow (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_p)$
    Call $SIMPLS(TrainX_{(-i)}, ClassY_{n\times g}, nfac)$ to get $Rd(X_{(-i)})$ , $Rd(Y_{(-i)})$
    $ieg(i) \leftarrow Rd(X)\text{-}Rd(X_{-i})$ ; (Definition 3)
    $deg(i) \leftarrow Rd(Y)\text{-}Rd(Y_{-i})$ ; (Definition 4)
Endfor
Return  $ieg$ , $deg$
```

```
Algorithm 3  $PLSVIP(TrainX_{n\times p}, ClassY_{n\times g}, nfac)$
Input:   $TrainX_{n\times p}, ClassY_{n\times g}, nfac$
Output:  $vip$

Begin
    Call $SIMPLS(TrainX_{n\times p}, ClassY_{n\times g}, nfac)$ to get calculated $Rd(X), Rd(Y), W$
    $Vip \leftarrow sqrt(p<Rd(Y), W^2>/Rd(Y))$   (Definition 5)
End
Return $vip$
```

Definition 3 (IEG: Independent-Variable Explanation Gain): Given the accumulation of variation explanation $Rd(X; t_1, t_2, \ldots, t_m)$ and $Rd_{(-i)}(X; t'_1, \ldots, t'_m)$ of component $t_1, t_2, \ldots, t_m$ and $t'_1, t'_2, \ldots, t'_m$ corresponding to independent variable X, we define the following:

$$\mathrm{IEG}(x_i) = Rd(X; t_1, \ldots, t_m) - Rd_{(-i)}(X; t'_1, \ldots, t'_m)$$

as independent variable explanation gain of x_i to independent variable *X*. Following the same logic, we have Definition 4.

Definition 4. DEG: Dependent-Variable Explanation Gain

Given the accumulation of variation explanation $Rd(Y; t_1, t_2, \ldots, t_m)$ and $Rd_{(-i)}(Y; t'_1, \ldots, t'_m)$ of component $t_1, t_2, \ldots, t_m$ and $t'_1, t'_2, \ldots, t'_m$ corresponding to dependent variable Y, we define the following:

$$\mathrm{DEG}(x_i) = Rd(Y; t_1, \ldots, t_m) - Rd_{(-i)}(Y; t'_1, \ldots, t'_m)$$

as the dependent variable explanation gain of x_i to dependent variable *Y.*

Larger value of $\mathrm{IEG}(x_i)$ indicates more importance of x_i in the interpretation of X. Similarly, the larger value of $DEG(x_i)$ indicates more importance of x_i in the interpretation of Y. Therefore, variable explanation gain $\mathrm{VEG}(x_i) = [\mathrm{IEG}(x_i), \mathrm{DEG}(x_i)]$ can be used to select features. Since Algorithm 2 calls Algorithm 1 repeatedly and the computational complexity of Algorithm 1 is $O(np)$, the computational complexity of Algorithm 2 approximates $O(np^2)$, a quadratic polynomial of p (gene number).

Definition 5 (VIP: Variant Importance in Projection [25]): Given the variation explanation of t_h to Y, and the accumulation of variation explanation $Rd(Y; t_h)$ and $Rd(Y; t_1, t_2, \ldots, t_m)$ of $t_1, t_2, \ldots, t_m$ to Y, we define the following:

$$\mathrm{VIP}(x_i) = \sqrt{\frac{p}{Rd(Y; t_1, \ldots, t_m)} \sum_{h=1}^{m} Rd(Y, t_h) w_{hi}^2}$$

as variant importance in projection of x_i to Y, where w_{hi} is the ith weight of axis w_h, which indicates the marginal contribution of x_i constructing components t_h.

To analyze the variable variation explanation with X to Y, we further quantitatively denote the impact of each x_i to Y. The variant importance in projection (VIP) was defined. The interpretation of x_i to Y is through t_h. When a strong explanatory power of t_h is to Y, and x_i plays an important role in structure t_h, the explanatory power of x_i to Y should be regarded as significant, i.e., if the values w_{hi} are also high based on the components t_h on large value $Rd(Y; t_h)$, then the interpretation of x_i to Y is strong. Larger $\mathrm{VIP}(x_i)$ indicates more importance in the interpretation of x_i to Y. Therefore, the indicator of VIP can be used in the feature selection. The computational complexity of Algorithm 3 is $O(np)$, a linear polynomial of p (gene number).

TABLE I
PLS-Based Feature-Selection Procedure

Step 1	*nfac*=0, *k*=0, *max_nfac*=g (number of category), *max_k*=100.
Step 2	Calculate the value of three *PLS*-based indicators (*VIP*, *IEG*, *DEG*) for each feature in the training set.
Step 3	Select the top *k* values on *PLS*-index in the training set for *SVMs* classification learning.
Step 4	Classify the testing set by *SVMs* using the top *k* selected features, and calculate the recognition rate.
Step 5	*k*=*k*+1, If *k* < *max_k,* goto Step3
Step 6	*nfac*=*nfac*+1; If *nfac* ≤ *max_nfac* ,repeat Step2 to Step5
Step 7	Maximize the classifier accuracy and at the same time minimize *k* from the *max_nfac* * *max_k* results.

IV. Experimental Results

The proposed PLS-based global gene-selection algorithm (Algorithms 2 and 3) is implemented in the MATLAB platform, and OSU_SVM3.00 Toolbox LinearSVC (parameter default) is used as classifiers. Since, we are focused on developing feature selection, it is feasible to use the widely applied classifier, i.e., SVM, as the baseline classifier to compare feature selection. In fact, an effective feature-selection algorithm is more important than the complexity of classifier.

First, we use the gene-selection Algorithms 2 and 3 presented earlier in this paper. The former k specific genes are selected from all the features in the training samples. Second, the classifier based on SVMs is trained by the selected specific genes. Finally, we test training samples and test samples. We attained the least number of specific genes, which identified correctly for all samples. The analysis results of the selected specific-gene subset are shown in the following. The steps of computing are shown in Table I.

In Table I, *nfac* is the number of factors in PLS model. More specifically, it is an input parameter for PLS-based feature selection. The feature-selection results depend on the value of *nfac* and the indicators. However, how to choose *nfac* is empirical. The value of *nfac* is usually less than or equal to the number of categories, since the purpose of the PLS model is to compress the original high-dimension feature genes into *nfac* potential genes in order to simplify the system and guide the classification. Regarding the tumor subtype classification problem on

TABLE II
SELECTED GENES BY DIFFERENT METHODS AND THE COMPARISONS WITH OTHER RESULTS

Data	*MIT AML/ALL(Leukemia Data)*				
Type	Experiment	Method	Classifier	Selected Gene	Identification
Two-Category	*Golub* [8]	*S2N*	*SVM*	50	(38,31)
	Golub [8]	*S2N*	*WV*	50	(36,29)
	Guyon [11]	*SVM+RFE*	*SVM*	8\16	(38,34)
	Li [17]	Model-free unbalanced sample	*SVM(Rbf)*	30	(38,34)
	Li [18]	*IIC*+Pair-wise redundancy+Sensitivity	*SVM(Rbf)*	16	(38,34)
	This paper	*PLSVIP*	*SVM(Linear)*	**9**	(38,34)
	This paper	*PLSIEG*	*SVM(Linear)*	**8**	(38,34)

TABLE III
SPECIFIC GENES SELECTED BY PLS

No.	*PLSVIP*		*PLSIEG*	
	VIP	Index No.	*IEG*	Index No.
1	0.000106	**6201**	4.74E-03	**6201**
2	9.88E-05	1674	4.44E-03	1674
3	8.51E-05	2186	3.20E-03	2186
4	8.03E-05	**1882**	2.43E-03	**1882**
5	7.29E-05	1376	2.02E-03	5976
6	7.22E-05	2402	1.87E-03	**6200**
7	6.80E-05	1394	1.84E-03	1394
8	6.56E-05	**6200**	1.73E-03	6806
9	6.18E-05	6803		

Data MIT AML/ALL(Leukemia) Data

two-category AML/ALL, it can be assumed that there are one or two latent genes that cannot be directly observed. Following the same logic, for the tumor subtype classification problem on four-category SRBCT, it can be assumed that there are no more than four latent genes that cannot be directly observed. The number of times the main loop (steps 3–5) iterates max_*k* times. Therefore, the time complexity for the aforementioned procedure is $O(np)$ using VIP indicator and $O(np^2)$ using IEG or DEG indicator.

A. Identification of Two-Category Tumor AML/ALL

The method proposed in this paper can achieve 100% recognition rate when 9 (8) genes are selected on data MIT AML/ALL (7129 genes). As to the PLS-based feature-selection method, *nfac* is set to 1 due to the fact that we can assume that there are one or two latent genes, which cannot be directly observed in the classification of tumor subtype on two-category AML/ALL. The linear kernel function is chosen in SVMs classifier, and default parameters are used. Table II shows the comparison.

Table II illustrates that the results obtained from VEG and VIP based on PLS are better than the results from Golub *et al.* [8]. Especially noteworthy, only eight feature genes can correctly identify all the training samples and test samples using PLSIEG method. The result is also superior to the other results in the literature. The eight feature genes are shown in Table III.

In the process of classification using LinearSVC, there are 38 training samples (ALL: 27; AML: 11), in which ALL only have four support vector (SVs) and AML only have three SVs. In the paper of Li and Ruan [18], a complex kernel function and kernel parameter (RbfSVC, sigma $= 10$, $C = 500$) was employed. Similar to Deng *et al.* [3], the level of significance (Sig.) is much less than 0.01 on the selected eight genes. To some extent, it is obvious that the PLS-based feature selection is more robust, and the generalization ability is stronger.

Nine specific genes and eight specific genes have been identified from all 7129 genes. Three of the top ten genes identified are (#6201, #1882, and #6200), which agrees with the literature [8], [33] and [34]. Table IV gives the description of some of the selected specific genes of MIT AML/ALL, and the visualization of three specific genes is shown in Fig. 1.

Fig. 1 is the scatter plot of the first three specific genes selected by PLS in MIT AML/ALL data set, which clearly shows the border of the two types (AML/ALL) with better separation. The information of sample types will improve the effectiveness of the follow-up supervised learning classification in PLS.

To visualize the differences before and after gene selection, we show the correlation matrix for the training samples on all genes and our selected genes. In each matrix, the (i, j) element is the value of the correlation coefficient between the ith and the jth samples in the training sets. The color intensity in those graphs reflects the magnitude of sample–sample correlation coefficient: the brighter the color is, the more similar the corresponding samples are.

Fig. 2 shows the graphs of correlation matrix before and after gene selection on the training samples (AML/ALL). Fig. 2(a) is the graph of correlation matrix of all the 7129 genes before gene selection. Fig. 2(b) is the graph of correlation matrix of eight specific genes selected by PLSIEG. Fig. 2(c) is the graph of correlation matrix of nine specific genes selected by PLSVIP. Fig. 2 shows a clear two-category contour in the training samples based on the specific genes chosen by PLS algorithm. Our proposed PLS gene selection can remove a large number of redundant genes. Therefore, it enhances the similarity of within-category samples on the selected genes, i.e., the information of classification between-category is clearer. It further approves the effectiveness of our methodology on removing redundant genes. It also shows that our feature-selection method is very effective. Fig. 3 shows the expression patterns in two different tumor subtypes (AML/ALL) on the specific genes chosen by PLS algorithm. It can be observed that the selected genes are significantly differently expressed in the training sample.

TABLE IV
DESCRIPTION OF THE SELECTED SPECIFIC GENE

No.	Index No.	Gene Description
1	6201(L19593_at)	IL8RB Interleukin 8 receptor, beta
2	1882(M27891_at)	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
3	6200(M28130_rna1_s_at)	Interleukin 8 (IL8) gene

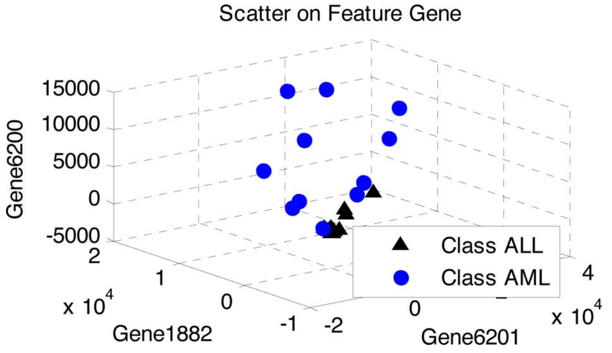


Fig. 1. Scatter plot of the first three specific genes on PLS (AML/ALL).

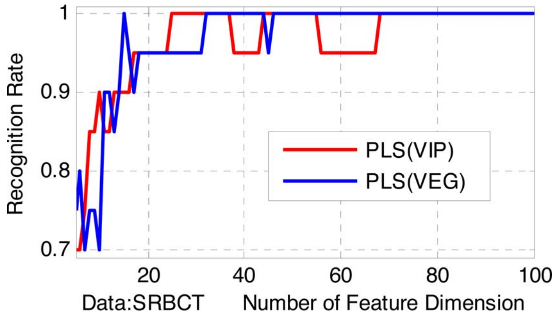


Fig. 4. Relationship between the number of feature dimension and recognition rate.

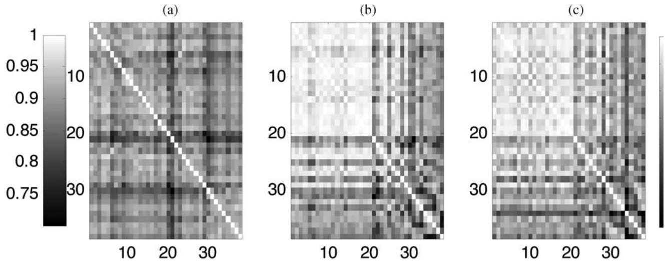


Fig. 2. Correlation matrix of training samples (AML/ALL). (a) All Genes. (b) Selected by PLSIEG. (c) Selected by PLSVIP.

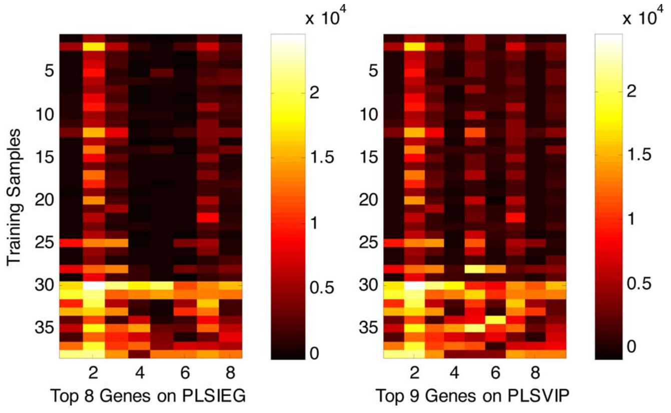


Fig. 3. Hot color map of expression intensity of top 8/9 specific genes on training samples (AML/ALL).

B. *Identification of Multicategory Tumor SRBCT*

Similar to SRBCT data, when *nfac* = 3 and 4, we can assume that there are no more than four latent genes, which cannot be directly observed in the classification of tumor subtype on four-category SRBCT.Twenty four and fifteen feature genes were selected respectively by VIP and VEG method based on PLS selection. Training samples and test samples were all identified correctly using linear kernel SVMs and default parameters. Table V shows the comparison.

Fig. 4 gives the relationship between the recognition rate on test samples and the top k specific genes based on VIP and VEG of PLS methods.

The smallest number of specific genes reaches 24 and 15 when recognition rate reaches 100%. It also indicates that the proposed method has a powerful classification ability and strong convergence in the multicategory classification problem. As can be seen from the recognition rate, the corresponding specific gene was superior to the literature when SRBCT data reached full recognition by VIP and VEG methods. Table VI gives the index number and the value of VIP and DEG selected on SRBCT.

Twenty-four specific genes and fifteen specific genes are identified from all 2308 genes, respectively. The top ten specific genes also included genes of #246, #545, #1389, and #1954. Up to six specific genes selected on DEG were consistent with the relevant studies [33], [34]. The result shows that the selection of specific gene on PLS was also reliable in the classification of multicategory (four categories). Table VII gives the description of some of the selected gene specific on SRBCT.

Similarly, to visualize the differences before and after gene selection, we show the correlation matrix for the training samples on all genes and our selected genes.

Similarly, for four-category classification data, Fig. 5 provides the graphs of correlation matrix before and after gene selection on the training samples (SRBCT). Fig. 5(a) is the graph of correlation matrix of all the 2308 genes before gene selection. Fig. 5(b) is the graph of correlation matrix of 15 specific genes selected by PLSDEG. Fig. 5(c) is the graph of correlation matrix of 24 specific genes selected by PLSVIP. Fig. 5 also clearly shows four-category contour in the training samples based on the specific genes chosen by PLS algorithm.

TABLE V
SELECTED GENES BY DIFFERENT METHODS AND THE COMPARISONS WITH OTHER RESULTS

Data	*SRBCT Data*				
Type	Experiment	Method	Classifier	Selected Gene	Identification
Multi-Category	*Kahn* [13]	*ANN*	*ANN*	96	(63,18)
	Tibshirani [27]	Nearest shrunken centroid	*NSC*	43	(63,20)
	Li [17]	Model-free unbalanced sample	*SVM(Rbf)*	74	(63,20)
	This paper	*PLSVIP*	*SVM(Linear)*	**24**	(63,20)
	This paper	*PLSVEG*	*SVM(Linear)*	**15**	(63,20)

TABLE VI
SELECTED THE MOST SIGNIFICANT SPECIFIC GENE ON PLS

No.	*PLSVIP*		*PLSDEG*	
	VIP	Index No.	*DEG*	Index No.
1	0.026741	509	0.002062	**246**
2	0.013817	187	0.001908	1955
3	0.006276	**246**	0.001740	**1389**
4	0.005333	**545**	0.001670	1834
5	0.004477	1915	0.001606	**545**
6	0.004356	1955	0.001553	187
7	0.004322	**1389**	0.001365	1781
8	0.004257	276	0.001193	509
9	0.004074	151	0.001149	**1954**
10	0.004019	**1954**	0.000966	1547
11	0.003691	1645	0.000832	**742**
12	0.003643	430	0.000828	735
13	0.003626	951	0.000828	1774
14	0.003531	**742**	0.000776	**1601**
15	0.003486	469	0.000755	1750
16	0.003304	1834		
17	0.003095	1093		
18	0.002797	1517		
19	0.002687	1319		
20	0.002658	**1372**		
21	0.002548	1295		
22	0.002432	1750		
23	0.002417	1708		
24	0.002267	1066		

Data *SRBCT*

TABLE VII
DESCRIPTION OF SPECIFIC-GENE FUNCTION SELECTED ON PLS

No.	Index No.	Gene Description
1	246	Caveolin 1,caveolae protein,22kD
2	545	Antigen identified by monoclonal antibodies 12E7,F21 and O13
3	1389	Fc fragment of IgG,receptor, transporter,alpga
4	1601	Microtubule-associated protein 1B
5	742	Transmembrane protein
6	1954	Follicular lymphoma variant translocation 1

It is obvious that our proposed PLS-based gene selection can remove a large number of redundant genes, and enhance the similarity of within-category samples on the selected genes. Thus, the information of classification between-category is clearer.

Fig. 6 shows the expression patterns in four different tumor subtypes (EWS/BL/NB/RMS) on the specific genes chosen by PLS algorithm. It can be observed that the selected genes are significantly differently expressed in the training sample.

C. More Performance Evaluation

The following is the further analysis based on PLSVIP. The other algorithms follow the same logic. Because there is correlation between genes, our gene selection considers all the genes.

Table VIII illustrates the ranking position of the top nine specific genes based on PLSVIP using other gene-selection methods. The specific gene (index #1376) is ranked fifth based on PLSVIP. However, it is ranked 221 by SNR-based gene selection, and 87 by t-test-based gene selection. It shows that the proposed gene selection can capture those specific–genes, which have relatively small main effect, but with strong interaction effect.

To further assess the robustness of the proposed gene-selection method, ten-fold cross validation (ten-fold CV) is adopted in the following experiments. All samples of the data sets are randomly divided into ten subsets, nine subsets of which are treated as training set, and the other subset as testing set. Repeat this process ten times. Thus, each subset has been used as a test set. At the same time, in order to avoid "selection bias," we select the features from the training sets, train the classifier (SVM) on the selected feature subset, make predictions on the test set, and record the recognition rate with ten-fold CV.

The recognition rate for the top k specific genes on the test set is shown in the following using the 10-fold CV. Assuming that the 100 random results follow normal distribution. The mean is estimated by the maximum likelihood, and 95% confidence interval is calculated. Visualization of the results is presented in Figs. 7 and 8.

Fig. 7 shows that the recognition rate of the proposed PLS-based gene selection is greater with small number of features selected in the data set (AML/ALL). The length of 95% confidence interval is less than 0.05, which indicates that the robustness of proposed PLS-based gene selection is very strong.

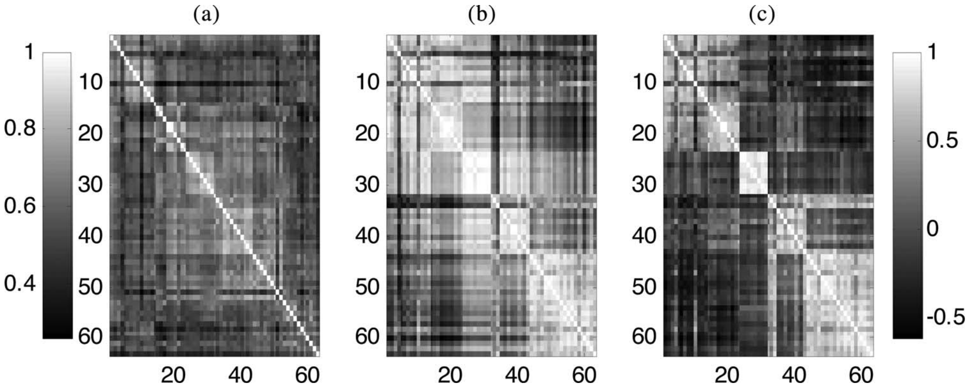


Fig. 5. Correlation matrix of training samples (SRBCT). (a) All Genes. (b) Selected by PLSDEG. (c) Selected by PLSVIP.

TABLE VIII
RANKING POSITION OF SELECTED GENE SPECIFIC IN THE OTHER GENE-SELECTION METHODS

Method		Index of the Selected specific-genes based on *PLSVIP*				
		6201	*1674*	*2186*	*1882*	1376
SNR	*SNR*	1.041	0.974	0.9225	1.1092	0.7171
	Rank	18	31	46	12	**221**
t-test	*Sig.*	1.76E-07	1.48E-06	1.66E-06	3.19E-07	7.55E-05
	Rank	12	24	26	15	**87**
Method		Index of the Selected specific-genes based on *PLSVIP*				
		2402	1394	6200	6803	
SNR	*SNR*	1.0054	0.8614	1.0009	0.9381	
	Rank	27	69	28	40	
t-test	*Sig.*	1.52E-06	4.93E-06	6.23E-07	1.43E-06	
	Rank	25	34	18	23	

Data MIT AML/ALL(Leukemia) Data

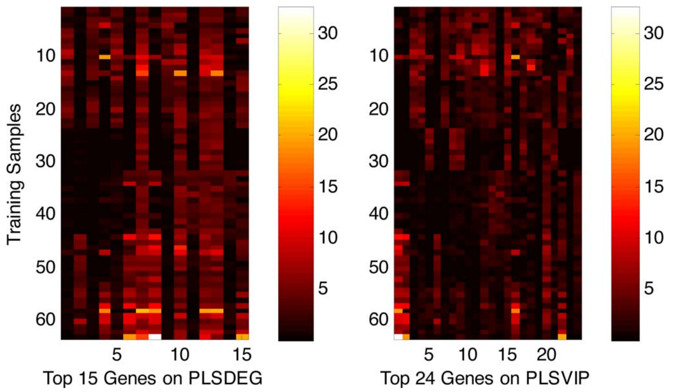


Fig. 6. Hot color map of expression intensity of top 15/24 specific genes on training samples (SRBCT).

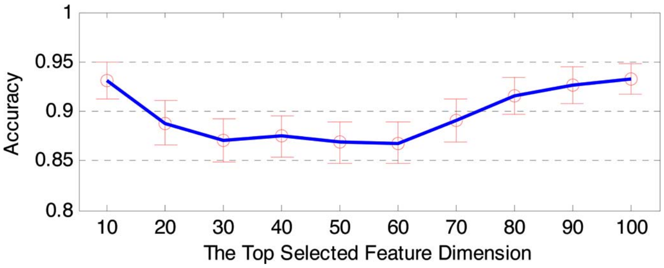


Fig. 7. Recognition rate of the top k specific genes by PLS-based gene selection (AML/ALL).

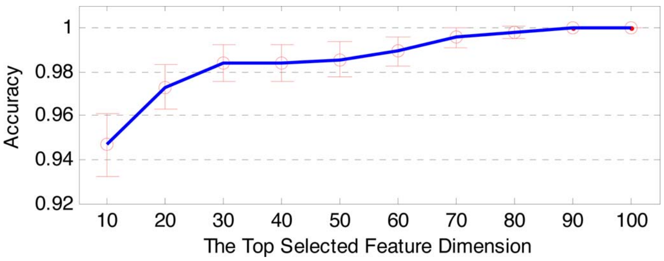


Fig. 8. Recognition rate of the top k specific genes by PLS-based gene selection (SRBCT).

It can be observed from Fig. 8 that the recognition rate increases when the number of specific genes chosen by the proposed PLS based increases in the data set (SRBCT). At the same time, the length of 95% confidence interval is getting smaller and smaller. The recognition rate is stable at 100%, when the number of specific genes chosen by our proposed algorithm increases to 85. It also shows that our gene selection is very robust. Moreover, the proposed method has been tested using other relatively new tumor data sets [5], [7], [15] (these data sets

TABLE IX
RECOGNITION RATE ON OTHER DATASETS USING PLSVIP-BASED GENE SELECTION WITH DIFFERENT CLASSIFIERS

Two-Category Data set	Classifier	Gene Selection			Gene Selection		
		PLSVIP(nfac=1)			PLSVIP(nfac=2)		
		Number	Training	Test	Number	Training	Test
Breast [5]	KNN(k=1)	**48**	**100%**	**89.47%**	52	100 %	89.47%
	KNN(k=3)	52	93.59%	84.21%	33	88.46%	84.21%
	SVM(Linear)	4	83.33%	78.95%	15	92.31%	89.47%
	SVM(Quadratic)	37	100%	84.21%	45	100%	89.47%
Lung [7]	KNN(k=1)	33	100%	100%	30	100%	100%
	KNN(k=3)	**25**	**100%**	**100%**	23	100%	100%
	SVM(Linear)	33	100%	100%	25	100%	99.33%
	SVM(Quadratic)	24	100%	100%	23	100%	100%
Multi-Category Data set		PLSVIP(nfac=5)			PLSVIP(nfac=7)		
		Number	Training	Test	Number	Training	Test
Leukemia [15] (Stjude data)	KNN(k=1)	64	100%	91.96%	38	100%	94.64%
	KNN(k=3)	91	94.42%	96.43%	36	91.63%	96.43%
	SVM(Linear)	**163**	**100%**	**97.32%**	136	100%	96.43%
	SVM(Quadratic)	101	100%	96.43%	146	100%	97.32%

are also small samples with high dimensions), and has shown the superior performance as well. See Table IX in Appendix II for details.

V. CONCLUSION

With high-dimension small sample of the tumor microarray data, we proposed the global gene-selection method by the definition of three PLS-based indicators (VIP, VEG, and IEG). The method considers the correlation between genes, and it can identify those genes, which have small main effects, but may have strong interaction effect. Tumor-specific genes are identified through the classification of the two tumors (MIT AML/ALL and SRBCT). Compared to the literature, the selection of specific genes on PLS has stronger recognition ability. Without setting complex parameters, our global gene-selection algorithm can efficiently select small numbers specific genes from a number of genes. Only a linear kernel SVM classifier is enough to reach 100% classification accuracy on the independent test samples.

This paper attempts to analyze the information of tumor gene expression in order to support the early diagnosis of cancer and guide the cancer prevention and treatment. As to these specific genes, we need to further investigate their auxiliary role in identification and the role of coregulation in the biochemical pathways, and in the occurrence and development of organisms, and explain the pathogenesis of cancer at the molecular level.

APPENDIX I

LIST OF ABBREVIATIONS

Term	Signification
SIMPLS	Simple partial least squares [4]
NIPALS	Nonlinear iterative partial least squares [28]
IEG	Independent-variable explanation gain
DEG	Dependent-variable explanation gain
VIP	Variant importance in projection
nfac	Number of factors in *PLS* model
$\boldsymbol{X} = (x_{ij})$	$n \times p$ matrix of predictor variables
$\boldsymbol{Y} = (y_{ij})$	$n \times q$ matrix response variables
$X_1, \ldots, X_p$	Centered to zero mean by each column
$Y_1, \ldots, Y_q$	Centered to zero mean by each column
$\boldsymbol{w}_j = (w_{1j}, \ldots, w_{pj})^T$	j-th latent component (weight vector for X)
$\boldsymbol{t}_j = (t_{1j}, \ldots, t_{nj})^T$	j-th latent component (score vector for X)
$\boldsymbol{c}_j = (c_{1j}, \ldots, c_{qj})^T$	j-th latent component (weight vector for Y)
$\boldsymbol{u}_j = (u_{1j}, \ldots, u_{nj})^T$	j-th latent component (score vector for Y)
$\boldsymbol{T} = [\boldsymbol{t}_1, \ldots, \boldsymbol{t}_{nfac}]$	$n \times$ *nfac* matrix of X scores
$\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{nfac}]$	$n \times$ *nfac* matrix of Y scores
$\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{nfac}]$	$p \times$ *nfac* matrix of weights
$\boldsymbol{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{nfac}]$	$p \times$ *nfac* matrix of X loadings
$\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{nfac}]$	$q \times$ *nfac* matrix of Y loadings
$\boldsymbol{E}$	$n \times p$ matrix of residual matrix for X
$\boldsymbol{F}$	$n \times q$ matrix of residual matrix for Y
$\boldsymbol{B}$	$p \times q$ matrix of regression coefficients

APPENDIX II

Because of the space limitation, we only provide the results of PLSVIP-based gene selection in the classifier KNN and SVM using three relatively new data sets. The results based on other indicators can be obtained easily following the same logic.

ACKNOWLEDGMENT

The authors would like to thank Prof. V. Marik and four anonymous reviewers for their constructive comments that improved the quality and the readability of this paper. The authors would also like to thank their laboratory team members' assistance.

REFERENCES

[1] M. Banerjee, S. Mitra, and H. Banka, "Evolutionary rough feature selection in gene expression data," *IEEE Trans. Syst., Man, Cybern. C*, vol. 37, no. 4, pp. 622–632, Jul. 2007.

[2] J. J. Dai, L. Lieu, and D. Rocke, "Dimension reduction for classification with gene expression microarray data," *Stat. Appl. Genet. Mol. Biol.*, vol. 5, no. 1, pp. 1–19, 2006.

[3] L. Deng, J. W. Ma, and J. Pei, "Rank sum method for related gene selection and its application to tumor diagnosis," *Chin. Sci. Bull.*, vol. 49, no. 15, pp. 1652–1657, 2004.

[4] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics Intell. Lab. Syst.*, vol. 18, no. 3, pp. 251–263, Mar. 1993.

[5] Y. Eng-Juh, E. R. Mary, A. S. Sheila, W. K. Williams, P. Divyen, M. Rami, G. B. Fred, C. R. Susana, V. R. Mary, P. Anami, C. Cheng, C. Dario, W. Dawn, X. D. Zhou, J. Y. Li, H. Q. Liu, C. H. Pui, E. E. William, N. Clayton, W. Limsoon, and R. D. James, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, pp. 133–143, Mar. 2002.

[6] J. F. Feng and J. X. Shi, "Gene selection based on fast fisher optimization model," *Acta Scientiarum Nat. Univ. Pekinensis*, vol. 41, no. 1, pp. 122–128, Jan. 2005.

[7] J. G. Gavin, V. J. Roderick, L. H. Li, R. G. Steven, E. B. Joshua, R. Sridhar, G. R. William, J. S. David, and B. Raphael, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Res.*, vol. 62, no. 1, pp. 4963–4967, Sep. 2002.

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

[9] B. F. Guo and M. S. Nixon, "Gait feature subset selection by mutual information," *IEEE Trans. Syst., Man, Cybern. A*, vol. 39, no. 1, pp. 36–46, Jan. 2009.

[10] M. Gutkin, G. Dror, and R. Shamir, "SlimPLS: A method for feature selection in gene expression-based disease classification," *PLoS One*, vol. 4, no. 7, p. pp. 1–12, e6416, 2009.

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

[12] B. Jose and B. A. Draper, "Feature selection from huge feature sets," in *Proc. 8th IEEE Conf. Comput. Vision Pattern Recognit.*, Jul. 2001, vol. 2, pp. 159–165.

[13] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nat. Med.*, vol. 7, no. 6, pp. 673–679, Jun. 2001.

[14] I. Kononenko, S. Robnik, and U. Pompe, "ReliefF for estimation and discretization of attributes in classification, regression and ILP problems," in *Proc. Artif. Intell. Methodol. Syst. Appl. (AIMSA) 1996*, IOS Press, pp. 31–40.

[15] J. V. Laura, D. Hongyue, J. V. Marc, D. H. Yudong, A. M. Augustinus, M. Mao, L. P. Hans, K. Karin, J. M. Matthew, T. W. Anke, J. S. George, M. K. Ron, R. Chris, S. L. Peter, B. Rene, and H. F. Stephen, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, Jan. 2002.

[16] G. Z. Li and X. Q. Zeng, "Feature selection for partial least square based dimension reduction," *Stud. Comput. Intell.*, vol. 205, pp. 3–37, 2009.

[17] J. Z. Li, K. Yang, H. Gao, J. Z. Luo, and Z. Guo, "Model-Free gene selection method by considering unbalanced samples," *J. Softw.*, vol. 17, no. 7, pp. 1485–1493, Jul. 2006.

[18] Y. X. Li and X. G. Ruan, "Feature selection for cancer classification based on support vector machine," *J. Comput. Res. Dev.*, vol. 42, no. 10, pp. 1796–1801, 2005.

[19] F. Li and Y. M. Yang, "Analysis of recursive gene selection approaches from microarray data," *Bioinformatics*, vol. 21, no. 19, pp. 3741–3747, Aug. 2005.

[20] A. Lorber, L. Wangen, and B. Kowalski, "A theoretical foundation for the PLS algorithm," *J. Chemometrics*, vol. 1, no. 1, pp. 19–31, 1987.

[21] S. Maldonado and R. Weber, "A wrapper method for feature selection using support vector machines," *Inf. Sci.*, vol. 179, no. 13, pp. 2208–2217, Jun. 2009.

[22] K. Z. Mao, "Feature subset selection for support vector machines through discriminative function pruning analysis," *IEEE Trans. Syst., Man, Cybern., B*, vol. 34, no. 1, pp. 60–67, Feb. 2004.

[23] D. V. Nguyen and D. M. Rocke, "Multi-class cancer classification via partial least squares with gene expression profiles," *Bioinformatics*, vol. 18, no. 9, pp. 1216–1226, 2002.

[24] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and minredundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[25] R. Rosipal and L. J. Trejo, "Kernel partial least squares regression in reproducing kernel Hilbert space," *J. Mach. Learn. Res.*, vol. 2, pp. 97–123, Mar. 2001.

[26] X. G. Ruan, Y. X. Li, J. G. Li, D. X. Gong, and J. L. Wang, "Tumor-specific gene expression patterns with gene expression profiles," *Sci. China Ser. C*, vol. 49, no. 3, pp. 293–304, 2006.

[27] R. Tibshirani, B. H. Narasimhan, and C. Gilbert, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proc. Nat. Acad. Sci.*, vol. 99, no. 10, pp. 6567–6572, Feb. 2002.

[28] H. Wold, "Path models with latent variables: The NIPALS approach," in *Quantitative Sociology: International Perspectives on Mathematical and Statistical Modeling*, H. M. Balock, A. Aganbegian, F. M. Borodkin, R. Boudon, and V. Cappecchi, Eds. New York: Academic Press, 1975, pp. 307–357.

[29] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, "The collinearity problem in linear regression, the partial least squares (PLS) approach to generalized inverses," *SIAM J. Sci. Stat. Comp.*, vol. 5, no. 3, pp. 735–743, Sep. 1984.

[30] E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," in *Proc. 18th Int. Conf. Mach. Learn. Morgan Kaufmann*, San Francisco, CA, 2001, pp. 601–608.

[31] X. Q. Zeng, G. Z. Li, G. F. Wu, J. Y. Yang, and M. Q. Yang, "Irrelevant gene elimination for partial least squares based dimension reduction by using feature probes," *Int. J. Data Mining Bioinformatics*, vol. 3, no. 1, pp. 85–103, Mar. 2008.

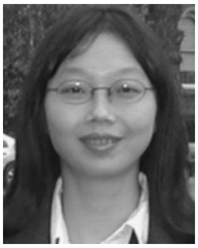
[32] X. Q. Zeng, G. Z. Li, M. Q. Yang, G. F. Wu, and J. Y. Yang, "Orthogonal projection weights in dimension reduction based on partial least squares," *Int. J. Comput. Intell. Bioinformatics Syst. Biol.*, vol. 1, no. 1, pp. 100–115, 2009.

[33] X. Zhou, X. Wang, and E. R. Dougherty, "Gene selection using logistic regression based on AIC, BIC and MDL criteria," *Biostatistics*, vol. 1, no. 1, pp. 129–145, 2005.

[34] X. Zhou, X. Wang, and E. R. Dougherty, "Nonlinear probit gene classification using mutual information and wavelet based feature selection," *J. Biol. Syst.*, vol. 12, no. 3, pp. 371–386, 2004.

Guoli Ji received the B.Sc. degree in automation control and the M.A.Sc. degree in system engineering from Xi'an Jiaotong University, Xi'an, P. R. China, in 1982 and 1986, respectively.

He is a Full Professor in the Department of Automation, Xiamen University, Xiamen, China. He is also the President of Institute of Systems Engineering, and the President of Xiamen Association of Systems Engineering. He has led and participated in many research projects from the Natural Science Foundation of China (NSFC), Natural Science Foundation of Fujian (NSFF), and others. Recently, he has published papers in BioMed Central (BMC) Bioinformatics, Nucleic Acids Research, BMC Genomics, Journal of Theoretical Biology, and Journal of Computational and Theoretical Nanoscience. His expertise and research areas include bioinformatics, computational biology and systems biology, advanced process control , model predictive control technology and software development, biological databases, data-mining technologies and platform development, modeling and simulation of complex systems, decision theory and decision support systems, management information systems and system integration for enterprises, etc.



Zijiang Yang received the M.A.Sc. and Ph.D. degrees in industrial engineering from the University of Toronto, ON, Canada, in 1999 and 2002, respectively.

She is currently an Associate Professor at the School of Information Technology, York University, Toronto, ON. Her current research interests include prediction, classification, performance analysis in the financial service industry, and data-mining algorithms. She is the author or coauthor of several published papers in IEEE Transactions on Engineering Management, IEEE Transactions on Neural Networks, Expert System with Applications, Computers and Operations Research, Journal of the Operational Research Society, Applied Mathematics and Computation, Mathematical and Computer Modeling, Annals of Operations Research, Communications in Nonlinear Science and Numerical Simulations, Socio-Economic Planning Sciences, and other peer-reviewed journals.

Wenjie You received the B.Sc. degree in mathematics from Zhangzhou Normal University, Zhangzhou, Fujian, China, in 1997, and M.Eng. degree in control engineering from Xiamen University, Xiamen, Fujian, in 2009, wherehe is currently working toward the Ph.D. degree in systems engineering.

Since 2003, he has been a Lecturer in the Department of Mathematics and Computer Science, Fujian Normal University, Fujian, China. His current research interests include statistical computing, data mining, and machine learning.

Mr. You has been a member of China Computer Federation since 2008.