Hindawi Publishing Corporation
Discrete Dynamics in Nature and Society
Volume 2015, Article ID 261013, 8 pages
http://dx.doi.org/10.1155/2015/261013

Research Article

Integrated Use of Statistical-Based Approaches and Computational Intelligence Techniques for Tumors Classification Using Microarray

Chia-Ding Hou and Yuehjen E. Shao

Department of Statistics and Information, Fu Jen Catholic University, New Taipei City 24205, Taiwan

Correspondence should be addressed to Yuehjen E. Shao; stat1003@mail.fju.edu.tw

Received 16 February 2015; Revised 1 April 2015; Accepted 9 April 2015

Academic Editor: Miguel Ángel López

With the recent development of biotechnologies, cDNA microarray chips are increasingly applied in cancer research. Microarray experiments can lead to a more thorough grasp of the molecular variations among tumors because they can allow the monitoring of expression levels in cells for thousands of genes simultaneously. Accordingly, how to successfully discriminate the classes of tumors using gene expression data is an urgent research issue and plays an important role in carcinogenesis. To refine the large dimension of the genes data and effectively classify tumor classes, this study proposes several hybrid discrimination procedures that combine the statistical-based techniques and computational intelligence approaches to discriminate the tumor classes. A real microarray data set was used to demonstrate the performance of the proposed approaches. In addition, the results of cross-validation experiments reveal that the proposed two-stage hybrid models are more efficient in discriminating the acute leukemia classes than the established single stage models.

1. Introduction

The recent development of cDNA microarray technologies has made it possible to analyze thousands of genes simultaneously and has led to the prospect of providing an accurate and efficient means for classifying and diagnosing human cancers [1–20]. Advances in microarray discrimination method promise to greatly advance cancer diagnosis, especially in situations where tumors are clinically atypical. The main challenge of microarray analysis, however, is the overwhelming number of genes compared to the smaller number of available tumor samples, that is, a very large number of variables relative to the number of observations [10, 21–23]. As a consequence, the issue of developing an accurate discrimination method for tumor classification using gene expression data has received considerable attention recently.

Many approaches have been proposed for tumor classification using microarray data [10, 22–33]. The existing methods can be divided into two types, the statistical-based methods [10, 22, 24–26] and computational intelligence methods [22, 27–33]. Due to the fact that the dimension of the genes data is very large, but there are only a few observations available, it is a must to reduce and refine the whole data set before we perform the classification tasks. While most related works have focused on the use of a single technique for tumor classification, little research has been done on the integrated use of several techniques simultaneously to classify tumor classes. To achieve the high accuracy for a particular classification problem with smaller computational time, hybrid evolutionary computation algorithms are commonly used for optimizing the resolution process [34–36]. As a consequence, in this study, we aim to develop several effective two-stage hybrid discrimination approaches that integrate the framework of statistical methods and the computational intelligence methods for tumors classification based on gene expression data.

The remainder of this paper is structured as follows. The second section reviews several existing approaches considered in our comparison study. The third section addresses the proposed hybrid approaches for tumors classification.

The fourth section shows classification results from the cross-validation. The final section reports the research findings and presents a conclusion to complete this study.

2. Review of Established Methods

Consider a two-class classification problem. Let $\underset{\sim}{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$ be the gene expression profile vector, where x_{ij} is the expression level of the jth gene in the ith tumor sample, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, r$. Let Y_i be a binary disease status variable (1 for case group π_1 and -1 for control group π_2 as a general example). Accordingly, the microarray data may be summarized as the following set:

$$\left\{\left(Y_1, \underset{\sim}{x}_1\right), \left(Y_2, \underset{\sim}{x}_2\right), \ldots, \left(Y_n, \underset{\sim}{x}_n\right)\right\}. \tag{1}$$

The following sections briefly review several well-known established microarray classification methods.

2.1. Fisher's Linear Discriminant Analysis. With the use of gene expression data, several studies proposed to apply Fisher's linear discriminant analysis (FLDA) to classify and diagnose cancer [10, 22, 24]. Assume that independent observation vectors $\underset{\sim}{z}_{11}, \underset{\sim}{z}_{12}, \ldots, \underset{\sim}{z}_{1n_1}$ and $\underset{\sim}{z}_{21}, \underset{\sim}{z}_{22}, \ldots, \underset{\sim}{z}_{2n_2}$ are obtained from the two known groups π_1 and π_2, respectively. Let

$$\begin{aligned}
\overline{\underset{\sim}{z}}_1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} \underset{\sim}{z}_{1i},\\
\overline{\underset{\sim}{z}}_2 &= \frac{1}{n_2}\sum_{i=1}^{n_2} \underset{\sim}{z}_{2i},\\
S &= \frac{1}{n}\left[\sum_{i=1}^{n_1}\left(\underset{\sim}{z}_{1i} - \overline{\underset{\sim}{z}}_1\right)\left(\underset{\sim}{z}_{1i} - \overline{\underset{\sim}{z}}_1\right)' \right.\\
&\quad \left. + \sum_{i=1}^{n_2}\left(\underset{\sim}{z}_{2i} - \overline{\underset{\sim}{z}}_2\right)\left(\underset{\sim}{z}_{2i} - \overline{\underset{\sim}{z}}_2\right)'\right],
\end{aligned} \tag{2}$$

where

$$n = n_1 + n_2 - 2. \tag{3}$$

To classify new observation $\underset{\sim}{z}_0$, we can utilize the following FLDA allocation rule:

$$\begin{aligned}
&\left(\overline{\underset{\sim}{z}}_1 - \overline{\underset{\sim}{z}}_2\right)' S^{-1}\underset{\sim}{z}_0 - \frac{1}{2}\left(\overline{\underset{\sim}{z}}_1 - \overline{\underset{\sim}{z}}_2\right)' S^{-1}\left(\overline{\underset{\sim}{z}}_1 + \overline{\underset{\sim}{z}}_2\right) \geq 0,\\
&\qquad\qquad \text{allocate } \underset{\sim}{z}_0 \text{ to group } \pi_1,\\
&\left(\overline{\underset{\sim}{z}}_1 - \overline{\underset{\sim}{z}}_2\right)' S^{-1}\underset{\sim}{z}_0 - \frac{1}{2}\left(\overline{\underset{\sim}{z}}_1 - \overline{\underset{\sim}{z}}_2\right)' S^{-1}\left(\overline{\underset{\sim}{z}}_1 + \overline{\underset{\sim}{z}}_2\right) < 0,\\
&\qquad\qquad \text{allocate } \underset{\sim}{z}_0 \text{ to group } \pi_2.
\end{aligned} \tag{4}$$

2.2. Logistic Regression. The microarray discrimination approach with the use of logistic regression (LR) model was also studied for disease classification [22, 25, 26]. The structure of the logistic regression model can be briefly described as follows. Let

$$P_i = \Pr\left[Y_i = 1 \mid x_{i1}, x_{i2}, \ldots, x_{ir}\right], \quad i = 1, 2, \ldots, n, \tag{5}$$

be the conditional probability of event $\{Y_i = 1\}$ under a given series of independent variables $(x_{i1}, x_{i2}, \ldots, x_{ir})$. The logistic regression model then is defined as follows:

$$\ln\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \sum_{j=1}^{r}\beta_j x_{ij}. \tag{6}$$

Collinearity diagnosis procedure should be conducted first to exclude variables exhibiting high collinearity. After collinearity diagnosis, the remaining variables are then used for logistic regression modeling and testing. Afterward, using logistic regression with Wald-forward method, we can identify significant independent variables, say, $x_{i1}^*, x_{i2}^*, \ldots, x_{ik}^*$, and obtain a significance model

$$\widehat{P}_i = \frac{\exp\left(\widehat{\beta}_0 + \sum_{j=1}^{k}\widehat{\beta}_j x_{ij}^*\right)}{1 + \exp\left(\widehat{\beta}_0 + \sum_{j=1}^{k}\widehat{\beta}_j x_{ij}^*\right)}, \quad i = 1, 2, \ldots, n. \tag{7}$$

2.3. Artificial Neural Network. Based on gene expression profiles, the artificial neural network (ANN) has also been used to discriminate the tumor classes [22, 27–29]. The ANN framework includes the input, the output, and the hidden layers. The nodes in the input layer receive input signals from an external source and the nodes in the output layer provide the target output signals. For each neuron j in the hidden layer and neuron k in the output layer, the net inputs are given by

$$\begin{aligned}
\text{net}_j &= \sum_i w_{ji} \times o_i,\\
\text{net}_k &= \sum_j w_{kj} \times o_j,
\end{aligned} \tag{8}$$

where i (j) is a neuron in the previous layer, w_{ji} (w_{kj}) is the connection weight from neuron i (j) to neuron j (k), and o_i (o_j) is the output of node i (j). The sigmoid functions are given by

$$\begin{aligned}
o_i &= \text{net}_i,\\
o_i &= \frac{1}{1 + \exp^{-(\text{net}_i + \theta_i)}} = f_i\left(\text{net}_i, \theta_i\right),\\
o_k &= \frac{1}{1 + \exp^{-(\text{net}_k + \theta_k)}} = f_k\left(\text{net}_k, \theta_k\right),
\end{aligned} \tag{9}$$

where net_j (net_k) is the input signal from the external source to the node j (k) in the input layer and θ_j (θ_k) is a bias. The conventional technique used to derive the connection weights of the feedforward network is the generalized delta rule [37].

2.4. Support Vector Machine. To classify tumor classes using microarray data, the discrimination method with the use of support vector machine (SVM) has also been discussed [22, 30–33]. The structure of SVM algorithm can be described as follows. Let $\{(y_i, \underset{\sim}{x}_i)\}_{i=1}^{n}$, $\underset{\sim}{x}_i \in R^r$, $y_i \in \{-1, 1\}$, be the training set with input vectors and labels, where n is the number of sample observations and r is the dimension of each observation, and y_i is known target. The algorithm is to seek the hyperplane $\underset{\sim}{w}' \cdot \underset{\sim}{x}_i + b = 0$, where $\underset{\sim}{w}$ is the vector of hyperplane and b is a bias term, to separate the data from two classes with maximal margin width $2/\|\underset{\sim}{w}\|^2$. In order to obtain the optimal hyperplane, the SVM was used to solve the following optimization problem:

$$\begin{aligned} \text{Min} \quad & \Phi\left(\underset{\sim}{x}\right) = \frac{1}{2}\left\|\underset{\sim}{w}\right\|^2 \\ \text{s.t.} \quad & y_i\left(\underset{\sim}{w}'\underset{\sim}{x}_i + b\right) \geq 1, \quad i = 1, 2, \ldots, n. \end{aligned} \tag{10}$$

Because it is difficult to solve (10), SVM transforms the optimization problem to be dual problem by Lagrange method. The value of α in the Lagrange method must be nonnegative real coefficients. Equation (10) is transformed into the following constrained form [38]:

$$\begin{aligned} \text{Max} \quad & \Phi\left(\widetilde{w}, b, \xi, \alpha, \beta\right) \\ & = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{N}\alpha_i\alpha_j y_i y_j \widetilde{x}_i^T \widetilde{x}_j \\ \text{s.t.} \quad & \sum_{j=1}^{n}\alpha_j y_j = 0; \quad 0 \leq \alpha_i \leq C, \ i = 1, 2, \ldots, n. \end{aligned} \tag{11}$$

In (11), C is the penalty factor and determines the degree of penalty assigned to an error. Typically, it could not find the linear separate hyperplane for all application data. For problems that can not be linearly separated in the input space, SVM employs the kernel method to transform the original input space into a high dimensional feature space, where an optimal linear separating hyperplane can be found. The common kernel functions are linear, polynomial, radial basis function (RBF), and sigmoid. Although several choices for the kernel function are available, the most widely used kernel function is the RBF which is defined as [39]

$$K\left(\widetilde{x}_i, \widetilde{x}_j\right) = \exp\left(-\gamma\left\|\widetilde{x}_i - \widetilde{x}_j\right\|^2\right), \quad \gamma \geq 0, \tag{12}$$

where γ denotes the width of the RBF. Consequently, the RBF is used in this study and the multiclass SVM method is used in this study [40].

2.5. Multivariate Adaptive Regression Splines. The multivariate adaptive regression splines (MARS) have also been applied for tumor classification using gene expression data [22, 30]. The general MARS function can be represented as follows:

$$\widehat{f}\left(x\right) = b_0 + \sum_{m=1}^{M} b_m \prod_{k=1}^{K_m}\left[S_{km}\left(x_{\nu(k,m)} - t_{km}\right)\right], \tag{13}$$

where b_0 and b_m are the parameters, M is the number of basis functions (BF), K_m is the number of knots, S_{km} takes on values of either 1 or −1 and indicates the right or left sense of the associated step function, $\nu(k, m)$ is the label of the independent variable, and t_{km} is the knot location. The optimal MARS model is chosen in a two-step procedure. Firstly, construct a large number of basis functions to fit the data initially. Secondly, basis functions are deleted in order of least contribution using the generalized cross-validation (GCV) criterion. To measure the importance of a variable, we can observe the decrease in the calculated GCV values when a variable is removed from the model. The GCV is defined as follows:

$$\text{LOF}\left(\widehat{f}_M\right) = \text{GCV}\left(M\right) = \frac{(1/n)\sum_{i=1}^{n}\left[y_i - \widehat{f}_M\left(x_i\right)\right]^2}{\left[1 - C\left(M\right)/n\right]^2}, \tag{14}$$

where n is the observations and $C(M)$ is the cost penalty measures of a model containing M basis function.

3. The Proposed Hybrid Discrimination Methods

The two-stage hybrid procedure is commonly used in various fields such as financial distress warning system [41, 42], medical area [43], statistical inference [44, 45], and statistical process control [36, 46–48]. To obtain the best accuracy for a specific classification problem, hybrid evolutionary computation algorithms are commonly used to optimize the resolution process [34–36]. In this section, several two-stage hybrid discrimination methods that integrate the framework of statistical-based approaches and computational intelligence methods are proposed for tumor classification based on gene expression microarray data.

The proposed methods include five components: the FLDA, the LR model, the MARS model, the ANN, and the SVM classifiers. The proposed hybrid discrimination methods combine the statistical-based discrimination methods and computational intelligence methods. In stage 1, influencing variables are selected using LR or MARS. In stage 2, the selected important influencing variables are then taken as the input variables of FLDA, LR, ANN, SVM, or MARS. The following sections address the proposed approaches.

3.1. Two-Stage Hybrid Method of LR and Various Classifiers

Stage 1. Substitute independent variables $x_{i1}, x_{i2}, \ldots, x_{ir}$ and dependent variable y_i into logistic regression. Apply logistic regression with Wald-forward method to identify significant independent variables, say, $x_{i1}^*, x_{i2}^*, \ldots, x_{ik}^*$.

Stage 2. Substitute the significant independent variables $x_{i1}^*, x_{i2}^*, \ldots, x_{ik}^*$ obtained in Stage 1 and dependent variable y_i into various classifiers such as FLDA, ANN, SVM, or MARS. The obtained corresponding hybrid methods are referred to as the LR-FLDA, LR-ANN, LR-SVM, and LR-MARS, respectively.

Table 1: The influencing genes selected by using two-sample t-test with a significance level of 0.0001.

Variables	Gene description		Mean	S.D.	P value
x_1	CMKBR7 chemokine (C-C) receptor 7	AML	68.12	145.55	0.00005
		ALL	−48.02	82.69	
x_2	LAMP2 lysosome-associated membrane protein 2 {alternative products}	AML	171.60	113.80	0.00001
		ALL	62.21	80.22	
x_3	Quiescin (Q6) mRNA, partial cds	AML	1534.92	1070.94	0.00006
		ALL	715.77	559.19	
x_4	Peptidyl-prolyl CIS-TRANS isomerase, mitochondrial precursor	AML	299.44	270.60	0.00006
		ALL	34.64	110.62	
x_5	Transmembrane protein mRNA	AML	90.04	72.82	0.00003
		ALL	18.98	59.12	
x_6	PGD phosphogluconate dehydrogenase	AML	970.52	621.00	0.00002
		ALL	480.91	313.41	
x_7	Canalicular multispecific organic anion transporter (cMOAT)	AML	42.56	63.05	0.00006
		ALL	131.85	84.13	
x_8	Huntingtin interacting protein (HIP1) mRNA	AML	−9.40	121.07	0.00000
		ALL	−136.09	120.05	
x_9	ME491 gene extracted from *H. sapiens* gene for Me491/CD63 antigen	AML	2026.80	1658.10	0.00001
		ALL	747.70	548.40	
x_{10}	GB DEF = nonmuscle myosin heavy chain-B (MYH10) mRNA, partial cds	AML	243.12	109.91	0.00002
		ALL	486.81	330.69	
x_{11}	P4HB procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), beta polypeptide	AML	2015.60	1384.83	0.00003
		ALL	1015.83	503.40	

3.2. Two-Stage Hybrid Method of MARS and Various Classifiers

Stage 1. Substitute independent variables $x_{i1}, x_{i2}, \ldots, x_{ir}$ and dependent variable y_i into multivariate adaptive regression splines. Use multivariate adaptive regression splines to identify significant independent variables, say, $x^*_{i1}, x^*_{i2}, \ldots, x^*_{ik}$.

Stage 2. Substitute the significant independent variables $x^*_{i1}, x^*_{i2}, \ldots, x^*_{ik}$ obtained in Stage 1 and dependent variable y_i into various classifiers such as FLDA, LR, ANN, or SVM. The corresponding hybrid methods are referred to as the MARS-FLDA, MARS-LR, MARS-ANN, and MARS-SVM, respectively.

4. The Cross-Validation Experiments

This study performs a series of cross-validation experiments to compare the proposed approaches with those previously discussed in literature. This study considers a leukemia dataset that was first described by Golub et al. [5] and was examined in Dudoit et al. [10] and Lee et al. [22]. This dataset contains 6817 human genes and was obtained from Affymetrix high-density oligonucleotide microarrays. The data consist of 25 cases of acute myeloid leukemia (AML) and 47 cases of acute lymphoblastic leukemia (ALL).

Since the dimension of the data is very large ($r = 6817$) but there are only a few observations ($n = 72$), it is essential to reduce and refine the whole set of genes (independent variables) before we can construct the discrimination model.

Table 2: Collinearity diagnosis for LR modeling.

Variables	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
VIF	2.09	2.97	2.76	2.56	1.61	3.46	1.33	1.88	6.85	1.58	5.47

To refine the set of genes, Golub et al. [5], Dudoit et al. [10], and Lee et al. [22] proposed the methods of subjective ratios to select genes. It is well known that the two-sample t-test is the most popular test to test for the differences between two groups in means. For the sake of strictness, instead of using a somewhat arbitrary criterion like that used in Golub et al. [5], Dudoit et al. [10], or Lee et al. [22], this study applies the two-sample t-test with a significance level of 0.0001 to select the influencing genes. The results are given in Table 1.

The significant variables selected using two-sample t-test are then served as the input variables of the established single stage discrimination methods reviewed in Section 2 and the proposed two-stage hybrid methods introduced in Section 3. To examine the presence of collinearity, the variance inflation factor (VIF) was calculated. As shown in Table 2, all the values of VIFs are less than 10. Consequently, there is no high collinearity among these variables. In addition, this study adopts the suggestions of Dudoit et al. [10] and Lee et al. [22] and performs a 2 : 1 cross-validation (training set : test set).

The difficulty with ANN is that the design parameters, such as the number of hidden layers and the number of neurons in each layer, have to be set before training process can proceed. User has to select the ANN structure and set the values of certain parameters for the ANN modeling process.

```
> #Find the best parameter gamma&cost
> p<-seq(-1,1,1)
> obj<-tune.svm(y~., data=train, sampling="cross", gamma=2^(p), cost=2^(p))
> obj
Parameter tuning of 'svm':
- sampling method: 10-fold cross validation
- best parameters:
gamma cost
   0.5    2
> #Building the SVM model
> svm.model<-svm(y~., data=train, type="C-classification", gamma=obj$best.parameters[[1]], cost=obj$best.parameters[[2]])
> #Classification capability: Train
> svm.pred<-predict(svm.model, train)
> tab<-table(predict=svm.pred, true=train[,1])
> tab
       true
predict  0  1
      0 17  0
      1  0 31
> cat('Accurate Classification Rate = ',100*sum(dig(tab))/sum(tab), '% \n')
Accurate Classification Rate = 100 %
> #Classification capability: Test
> svm.pred<-predict(svm.model, test)
> tab<-table(predict=svm.pred, true=test[,1])
> tab
       true
predict  0  1
      0  2  1
      1  6 15
> cat('Accurate Classification Rate = ',100*sum(dig(tab))/sum(tab), '% \n')
Accurate Classification Rate = 70.83333 %
```

Algorithm 1: The SVM modeling output.

However, there is no general and explicit approach to select optimal parameters for the ANN models [49]. Accordingly, the selection of design parameters for ANN may be based on the trial and error procedure.

This study employs the highest accurate classification rate (ACR) as the criterion for selecting the ANN topology. The topology is defined as $\{n_i\text{-}n_h\text{-}n_o\text{-}L\}$, where it stands for the number of neurons in the input layer, number of neurons in the hidden layer, number of neurons in the output layer, and learning rate, respectively. Actually, too few hidden nodes would limit the network generation capability, while too many hidden nodes may result in overtraining or memorization by the network. Since there are 11 input nodes and one output node used in this study, the numbers of hidden nodes to test were selected as 9, 10, 11, 12, and 13. The learning rates are chosen as 0.1, 0.01, or 0.001, respectively. After performing the ANN modeling, this study found that the {11-9-1-0.01} topology has the best ACR results.

This study also performed the SVM modeling to the microarray dataset. The two parameters, C and γ, are the most important factors to affect the performance of SVM. The grid search method uses exponentially growing sequences of C and γ to determine good parameters. The parameter set of C and γ which generates the highest ACR is considered to be ideal set. Here, the best two parameter values for C and γ are 2 and 0.5, respectively. The SVM package was performed in running the dataset, and the corresponding output is displayed in Algorithm 1. Observing Algorithm 1, in the case of $C = 2$ and $\gamma = 0.5$, we can have ACR = 100% for the initial training stage. Consequently, in the testing stage, we are able to obtain ACR = 25% and ACR = 93.75% for AML and ALL, respectively, by using the same parameter settings (i.e., $C = 2$ and $\gamma = 0.5$). Accordingly, the ACR = 70.83% for the case of full sample.

Table 3: The relative importance of four explanatory variables for MARS modelling.

Function	Variable	Relative importance (%)
1	x_2	100.0
2	x_7	72.0
3	x_8	42.7
4	x_6	26.3

For MARS modeling, the results are displayed in Table 3. During the selection process, four important explanatory variables were chosen. The corresponding relative importance indicators are showed in Table 3. As a consequence,

Table 4: ACRs for thirteen approaches using cross-validation.

Method	ACR		
	AML	ALL	Full sample
Single stage			
FLDA	37.50%	93.75%	75.00%
LR	62.50%	87.50%	79.17%
ANN	50.00%	93.75%	79.17%
SVM	25.00%	93.75%	70.83%
MARS	50.00%	75.00%	66.67%
Two-stage			
LR-FLDA	62.50%	81.25%	75.00%
LR-ANN	50.00%	93.75%	79.17%
LR-SVM	75.00%	81.25%	79.17%
LR-MARS	62.50%	93.75%	83.33%
MARS-FLDA	75.00%	75.00%	75.00%
MARS-LR	75.00%	75.00%	75.00%
MARS-ANN	37.50%	93.75%	75.00%
MARS-SVM	62.50%	87.50%	79.17%

those four important variables would be served as the input variables for hybrid modeling process. In addition, the results of ACR for each modeling are listed in Table 4.

The rationale behind the proposed hybrid discrimination method is to obtain the fewer but more informative variables by performing the first stage LR or MARS modeling. The selected significant variables are then served as the inputs for the second stage of discrimination approach. In this study, the significant variables selected by performing LR and MARS modeling are variables x_1, x_2, x_7, and x_8 and variables x_2, x_6, x_7, and x_8, respectively. For the hybrid LR-ANN model, the {4-6-1-0.01} topology provided the best ACR results. For the MARS-ANN hybrid model, the {4-6-1-0.01} topology also gave the best ACR results. Additionally, for both LR-SVM and MARS-SVM modeling, the best two parameter values for C and γ are the same and they are 2 and 0.5, respectively.

For each of the thirteen different approaches, FLDA, LR, ANN, SVM, MARS, LR-FLDA, LR-ANN, LR-SVM, LR-MARS, MARS-FLDA, MARS-LR, MARS-ANN, and MARS-SVM, this study presents the corresponding ACRs in Table 4. By comparing the ACR with AML, while the LR has highest ACR (i.e., 62.50%) among the 5 single stage methods, both LR-SVM and MARS-LR have the highest ACR (i.e., 75.00%) among the 8 two-stage methods. Apparently, the two-stage methods provide a better classification performance. By comparing the ACR with ALL, the single stage methods of FLDA, ANN, and SVM give the highest ACR (i.e., 93.75%), and the two-stage methods of LR-ANN, LR-MARS, and MARS-ANN have the same ACR (i.e., 93.75%). It seems that the single stage and two-stage methods achieve a similar performance. As shown in Table 4, it can be seen that, among the thirteen methods mentioned above, the two-stage hybrid model of LR-MARS has the highest ACRs (i.e., 83.33%) for the full sample. As a consequence, the proposed two-stage hybrid approaches are more efficient for tumor classification than the established single stage methods.

Table 5: Overall averaged ACR and the associated standard error (in parentheses) for single stage and two-stage methods.

Method	ACR		
	AML	ALL	Full sample
Single stage	45.00% (14.25%)	88.75% (8.15%)	74.17% (5.43%)
Two-stage	62.50% (13.36%)	85.16% (8.14%)	77.61% (3.10%)

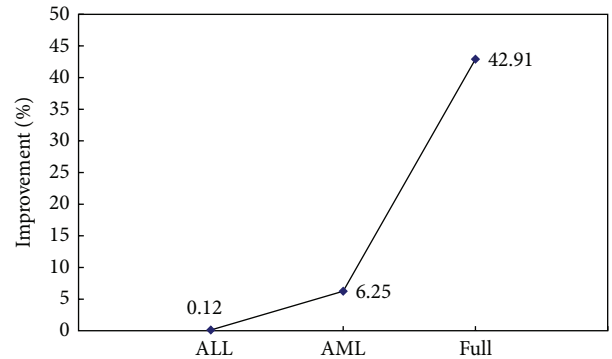

Figure 1: Improvement of the proposed approach in comparison with the single stage method.

In addition, Table 5 lists the overall averaged ACRs and the associated standard errors (in parentheses) for single stage and two-stage methods. In comparison to the single stage and the proposed two-stage methods in Table 5, one is able to observe that our proposed methods almost provide more accurate results than the single stage methods. Although the single stage methods have larger averaged ACR value than two-stage methods in classifying ALL, the difference is not too significant. In addition, observing Table 5 it can be found that the proposed two-stage approaches have the smaller standard errors for all the cases, which imply the robustness of the mechanisms. Figure 1 provides a comparison with respect to the overall improvement percentage in the single stage method. From Figure 1, it can be seen that the two-stage approaches are more robust than the single stage method.

5. Conclusions

This study proposes several two-stage hybrid discrimination approaches for tumor classification using microarray data. The proposed approaches integrate the framework of several frequently used statistical-based discrimination methods and computational intelligence classifying techniques. Based on the results of cross-validation in Table 4, it can be easily observed that the proposed hybrid method LR-MARS is more appropriate for discriminating the tumor classes.

Computational intelligence methodology is very useful in many aspects of application and can deal with complex and computationally intensive problems. With the use of several computational intelligence techniques, this study develops

two-stage hybrid discrimination approach for tumor classification. The proposed hybrid model is not the only discrimination method that can be employed. Based on our work further research can be expanded. For example, one can combine other computational intelligence techniques, such as rough set theory [50] or extreme learning machine, with neural networks or support vector machine to refine the structure further and improve the classification accuracy. Extensions of the proposed two-stage hybrid discrimination method to other statistical techniques or to multistage discrimination procedures are also possible. Such works deserve further research and are our future concern.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work is partially supported by the Ministry of Science and Technology of China, Grant no. MOST 103-2118-M-030-001 and Grant no. MOST 103-2221-E-030-021.

References

[1] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, 1997.

[2] R. J. Cho, M. J. Campbell, E. A. Winzeler et al., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, no. 1, pp. 65–73, 1998.

[3] S. Chu, J. DeRisi, M. Eisen et al., "The transcriptional program of sporulation in budding yeast," *Science*, vol. 282, no. 5389, pp. 699–705, 1998.

[4] U. Alon, N. Barka, 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 of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999.

[5] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[6] C. M. Perou, S. S. Jeffrey, M. van de Rijn et al., "Distinctive gene expression patterns in human mammary epithelial cells and breast cancers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 16, pp. 9212–9217, 1999.

[7] J. R. Pollack, C. M. Perou, A. A. Alizadeh et al., "Genome-wide analysis of DNA copy-number changes using cDNA microarrays," *Nature Genetics*, vol. 23, no. 1, pp. 41–46, 1999.

[8] A. A. Alizadeh, M. B. Elsen, R. E. Davis et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[9] S. Ramaswamy, P. Tamayo, R. Rifkin et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 15149–15154, 2001.

[10] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.

[11] J. J. Liu, G. Cutler, W. Li et al., "Multiclass cancer classification and biomarker discovery using GA-based algorithms," *Bioinformatics*, vol. 21, no. 11, pp. 2691–2697, 2005.

[12] L. Ziaei, A. R. Mehri, and M. Salehi, "Application of artificial neural networks in cancer classification and diagnosis prediction of a subtype of lymphoma based on gene expression profile," *Journal of Research in Medical Sciences*, vol. 11, no. 1, pp. 13–17, 2006.

[13] Z. Wang, Y. Wang, J. Xuan et al., "Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data," *Bioinformatics*, vol. 22, no. 6, pp. 755–761, 2006.

[14] K. V. G. Rao, P. P. Chand, and M. V. R. Murthy, "A neural network approach in medical decision systems," *Journal of Theoretical and Applied Information Technology*, vol. 3, pp. 97–101, 2007.

[15] H. Rattikorn and K. Phongphun, "Tumor classification ranking from microarray data," *BMC Genomics*, vol. 9, no. 2, article S21, 2008.

[16] L.-T. Huang, "An integrated method for cancer classification and rule extraction from microarray data," *Journal of Biomedical Science*, vol. 16, article 25, 10 pages, 2009.

[17] T. Y. Yang, "Efficient multi-class cancer diagnosis algorithm, using a global similarity pattern," *Computational Statistics & Data Analysis*, vol. 53, no. 3, pp. 756–765, 2009.

[18] H. Pang, K. Ebisu, E. Watanabe, L. Y. Sue, and T. Tong, "Analysing breast cancer microarrays from African Americans using shrinkage-based discriminant analysis," *Human Genomics*, vol. 5, no. 1, pp. 5–16, 2010.

[19] N. B. Dawany, W. N. Dampier, and A. Tozeren, "Large-scale integration of microarray data reveals genes and pathways common to multiple cancer types," *International Journal of Cancer*, vol. 128, no. 12, pp. 2881–2891, 2011.

[20] R. Pillai, R. Deeter, C. T. Rigl et al., "Validation and reproducibility of a microarray-based gene expression test for tumor identification in formalin-fixed, paraffin-embedded specimens," *The Journal of Molecular Diagnostics*, vol. 13, no. 1, pp. 48–56, 2011.

[21] D. Ghosh, "Penalized discriminant methods for the classification of tumors from gene expression data," *Biometrics*, vol. 59, no. 4, pp. 992–1000, 2003.

[22] J. W. Lee, J. B. Lee, M. Park, and S. H. Song, "An extensive comparison of recent classification tools applied to microarray data," *Computational Statistics & Data Analysis*, vol. 48, no. 4, pp. 869–885, 2005.

[23] M. S. Srivastava and T. Kubokawa, "Comparison of discrimination methods for high dimensional data," *Journal of the Japan Statistical Society*, vol. 37, no. 1, pp. 123–134, 2007.

[24] R. Bermudo, D. Abia, A. Mozos et al., "Highly sensitive molecular diagnosis of prostate cancer using surplus material washed off from biopsy needles," *British Journal of Cancer*, vol. 105, no. 10, pp. 1600–1607, 2011.

[25] W. Li, F. Sun, and I. Grosse, "Extreme value distribution based gene selection criteria for discriminant microarray data analysis using logistic regression," *Journal of Computational Biology*, vol. 11, no. 2-3, pp. 215–226, 2004.

[26] J. G. Liao and K. V. Chin, "Logistic regression for disease classification using microarray data: model selection in a large *p* and small *n* case," *Bioinformatics*, vol. 23, no. 15, pp. 1945–1951, 2007.

[27] S. Gruvberger, M. Ringnér, Y. Chen et al., "Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns," *Cancer Research*, vol. 61, no. 16, pp. 5979–5984, 2001.

[28] J. Khan, J. S. Wei, M. Ringnér et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[29] L. J. Lancashire, C. Lemetre, and G. R. Ball, "An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies," *Briefings in Bioinformatics*, vol. 10, no. 3, pp. 315–329, 2009.

[30] M. R. Segal, K. D. Dahlquist, and B. R. Conklin, "Regression approaches for microarray data analysis," *Journal of Computational Biology*, vol. 10, no. 6, pp. 961–980, 2003.

[31] A. Dragomir and A. Bezerianos, "Improving gene expression sample classification using support vector machine ensembles aggregated by boosting," *Cancer Genomics & Proteomics*, vol. 3, no. 1, pp. 63–70, 2006.

[32] R. Zhang, G.-B. Huang, N. Sundararajan, and P. Saratchandran, "Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 485–494, 2007.

[33] X. Wang and R. Simon, "Microarray-based cancer prediction using single genes," *BMC Bioinformatics*, vol. 12, article 391, 2011.

[34] K. Y. Chan, C. K. Kwong, and Y. C. Tsim, "Modelling and optimization of fluid dispensing for electronic packaging using neural fuzzy networks and genetic algorithms," *Engineering Applications of Artificial Intelligence*, vol. 23, no. 1, pp. 18–26, 2010.

[35] K. Y. Chan, T. S. Dillon, and C. K. Kwong, "Modeling of a liquid epoxy molding process using a particle swarm optimization-based fuzzy regression approach," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 1, pp. 148–158, 2011.

[36] Y. E. Shao and C.-D. Hou, "Change point determination for a multivariate process using a two-stage hybrid scheme," *Applied Soft Computing*, vol. 13, no. 3, pp. 1520–1527, 2013.

[37] D. E. Rumelhart and J. L. McClelland, *Explorations in the Microstructure of Cognition 1*, MIT Press, 1986.

[38] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, Berlin, Germany, 2000.

[39] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Networks*, vol. 17, no. 1, pp. 113–126, 2004.

[40] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[41] S. L. Lin, C. D. Hou, and P. H. Gi, "Do the two-stage hybrid models outperform the conventional techniques? Evidence in Taiwan," *International Journal of Business and Strategy*, vol. 9, pp. 98–131, 2008.

[42] S. L. Lin, "A new two-stage hybrid approach of credit risk in banking industry," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8333–8341, 2009.

[43] S.-M. Chou, T.-S. Lee, Y. E. Shao, and I.-F. Chen, "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines," *Expert Systems with Applications*, vol. 27, no. 1, pp. 133–142, 2004.

[44] R. Modarres and J. L. Gastwirth, "Hybrid test for the hypothesis of symmetry," *Journal of Applied Statistics*, vol. 25, no. 6, pp. 777–783, 1998.

[45] R. Tang, M. Banerjee, and G. Michailidis, "A two-stage hybrid procedure for estimating an inverse regression function," *The Annals of Statistics*, vol. 39, no. 2, pp. 956–989, 2011.

[46] W. Bischoff and F. Miller, "A minimax two-stage procedure for comparing treatments: looking at a hybrid test and estimation problem as a whole," *Statistica Sinica*, vol. 12, no. 4, pp. 1133–1144, 2002.

[47] C.-J. Lu, Y. E. Shao, and P.-H. Li, "Mixture control chart patterns recognition using independent component analysis and support vector machine," *Neurocomputing*, vol. 74, no. 11, pp. 1908–1914, 2011.

[48] Y. E. Shao and C. D. Hou, "Fault identification in industrial processes using an integrated approach of neural network and analysis of variance," *Mathematical Problems in Engineering*, vol. 2013, Article ID 516760, 7 pages, 2013.

[49] Z. W. Zhong, L. P. Khoo, and S. T. Han, "Prediction of surface roughness of turned surfaces using neural networks," *International Journal of Advanced Manufacturing Technology*, vol. 28, no. 7-8, pp. 688–693, 2006.

[50] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modeling schemes for heart disease classification," *Applied Soft Computing Journal*, vol. 14, pp. 47–52, 2014.