A GA Based Approach to Improving the ICA Based Classification Models for Tumor Classification

Jian-Bing Xia-Hou
School of Software
Xiamen University
Xiamen, Fujian Province
China
jbxiahou@xmu.edu.cn

Kun-Hong Liu* (Corresponding Author)
School of Software
Xiamen University
Xiamen, Fujian Province
China
lkhqz@xmu.edu.cn

Abstract: - As it has been pointed out that different ICs are of different biological significance, this paper tries to explore the IC selection problem based on a set of experiments. A regression model and a classification model, referred as penalized independent component regression (P-ICR) and ICA based Support Vector Machine (ICA+SVM), are applied to illustrate the necessity and efficiency of IC selection. A genetic algorithm (GA) is deployed to the selection process, along with an early stopping technique deployed to avoid overfitting in evolution. In particular, the individuals in the selected generation are used to construct an ensemble system to achieve higher classification accuracy. We test the two models with and without the selection methods based on three microarray datasets. The experiment results demonstrate that IC selection methods can further improve the classification accuracy of the ICA based prediction models, and the GA is more effective than the original methods.

Key-Words: - Microarray data; Independent component analysis (ICA); Genetic algorithm (GA); Early stopping; Support Vector Machine(SVM); Overfitting;

1 Introduction

With the development of microarray technology, it is possible to diagnose and classify some particular cancers directly based on DNA microarray data. However, the class prediction of microarray data is a typical "large p, small n" problem [1], which means that the number of predictor variables is much larger than the number of samples. So it is a great challenge to develop a new efficient method for analyzing global gene expression data. Currently a variety of algorithms and mathematical models have been used for management, analysis and interpretation of these high-density microarray data.

Principal component analysis (PCA) is a widely deployed tool for finding useful eigenassay or eigengene. With its aid, the gene snapshot coordinates can be predicted from each other linearly. But independent component analysis (ICA) has the great potential advantages over PCA in many aspects [2]. As a result, ICA technique has been attracting more and more attentions so far.

To illustrate a typical ICA processing, assume that an $n\times p$ data matrix X, with rows r_i (i=1,…,n) corresponding to observational variables and columns c_j (j=1,…,p), is the individual of the corresponding variables, the ICA model of X can be written as:

$$X=AS \quad (1)$$

Without loss of generality, A is an $n\times n$ matrix, and S is an $n\times p$ source matrix. Those variables in the rows of S are ICs, and the statistical independence between variables can be quantified by mutual information $I=\sum H(S_k)-H(S)$, where $H(S_k)$ is the marginal entropy of the variable S_k, and $H(S)$ is the joint entropy. And estimating the independent components can be accomplished according to the formula:

$$U=S=A^{-1}X=WX \quad (2)$$

Let matrix X denote the gene expression data, then it can be described as a linear mixture of statistically independent basis snapshots (eigenassay) S combined by an unknown mixing matrix A. In this approach, ICA is used to find a matrix W such that the rows of U are as statistically independent as possible. The representation of snapshots consists of their corresponding coordinates with respect to the eigenassays defined by the rows of U, i.e.

$$\mathbf{r}_j = a_{j1}\mathbf{u}_1 + a_{j2}\mathbf{u}_2 + \cdots + a_{jn}\mathbf{u}_n \quad (3)$$

The ICA transformation has been applied to the analysis of microarray data with great success. For example, Libermeister applied ICA to gene expression data and derived a linear model based on hidden variables [3], Lee and Batzoglou projected microarray data into statistically independent components and found that ICA outperformed other learning algorithms [4], and Zhang et al. extracted a set of specific diagnostic patterns of normal and tumor tissues corresponding to a set of biomarkers for clinical use based on ICA [5].

There is still not enough investigation on the ICA component selection problem, but it has been pointed out that dominant independent components could be related to particular biological or experimental effects, and the component weights (A) could be either tumor cluster or chromosomal aberration specific [3, 6]. So IC selection is an efficient technique to improve the performance of ICA based prediction system. In [6], the authors suggested that the most reliable way to select components is to manually inspect and evaluate them according to the corresponding component loading. But it is obviously impractical to identify the biological significance of each IC each time. And it is important for us to note that the best components do not necessarily constitute the best subset for prediction, and a given component will give out more information when presented with certain other components than considered only by it [7]. So a best IC subset for a prediction system should not contain the ICs with most biological significance. Instead, some seemingly less important ICs would give an aid to boost the final prediction accuracy greatly.

In order to select a proper subset of ICs for prediction, it is important to decide which ICs and how many ICs should be selected. Unfortunately, unlike PCA, there is not a universal rule for IC selection. The reasons lie in some aspects. First, the energies of the independent components cannot be determined immediately. Second, ICA is not always reproducible when used to analyze gene expression data [3], so a preset rule for analyzing the results of an algorithm may not always be applicable. Third, the results obtained from an ICA algorithm are not "ordered", and different source matrices can be generated by setting different number of ICs for a same observed signal. Fourth, different IC sets could be generated by different ICA algorithms for even a same source data. Based on the observations, we find that it is impossible to set up a simple and universal rule to guide the IC selection problem. Instead, the application of feature selection algorithms is a promising solution. And up to now, some authors have successfully applied different feature selection algorithms to deal with IC selection problem in different classification tasks, including EEG signal, face recognition, iris recognition [8-12]. And the experimental results showed that when applied in the IC selection problem, sequential floating forward selection (SFFS) outperforms the other algorithms[11]. But it is pointed out that usually GA is more efficient in feature selection [13].

In [31], a standard GA was designed to select ICs from a IC set, and then the IC subsets were used to train base classifiers so as to build an ensemble system. Furthermore, we found that due to the high dimension in microarray data, as pointed out above, different IC sets could be generated after different runs of the FastICA[17] algorithm on the same data. So in [32-33], for each microarray data, different IC sets were produced by the FastICA algorithm firstly, then a multi-objective GA was deployed to select different IC subsets from different IC sets. In this way, by using the IC subsets to train base classifiers, the diversity of base classifier was improved, so that the ensemble systems would work more effectively. In these papers, the selection of IC sets was all based on the last generation of the GAs, and the overfitting problem was not discussed.

In this paper, the early-stopping technique, which was designed to stop the training process of neural network, is deployed to determine in which generation the overfitting occurs. And then, the classifiers produced in a selected generation of the GA are combined to form an ensemble system. In this way, we can improve the performance of IC selection algorithms. So this algorithm is different from the GAs proposed in [10, 12, 31-33]. To validate the GA's performance, SFFS is also applied for comparison here. We discuss the IC selection problem empirically. In previous study, the benefits of IC selection have not been studied in different models for comparisons. Here, the ICA with SVM model (ICA+SVM) [14] and penalized independent component regression (P-ICR) model [15] are used verify the importance of IC selection. It should be

noted that these two models are set up according to completely different principles: the first one is a classification model, and the latter is a regression model. By addressing the importance of IC selection based on two different models, the discussions are of general sense and can be easily extended to other ICA based models. So the conclusions would be helpful when designing or using other ICA based prediction systems.

This paper is organized as follows. Section 2 presents the two ICA based models and the design scheme of the GA. In Section 3, a set of experimental results is presented to demonstrate the effectiveness of the IC selection method along with corresponding discussions. This paper is concluded in Section 4.

2 Reasearch Methodology

2.1 Two ICA based prediction models

Two completely different models are deployed in this study: P-ICR model and ICA+SVM. The former is a regularized regression model, with the aid of the optimal scoring algorithm [16] for the implement of classification. While the later is a classification model. Due to the different prediction principles, these two models are representative of different ICA based models. Below the frameworks of the two models are given roughly for the integrality purpose.

(a) The P-ICR model

P-ICR model is based on a regression method. Firstly, a singular value decomposition (SVD) is performed on the gene data $\mathbf{X}$:

$$\mathbf{X}=\mathbf{UDV}^T \tag{4}$$

Here, U is an n×n singular value decomposition matrix. The diagonal matrix $\mathbf{D}$ contains the ordered eigenvalues of $\mathbf{XX}^T$, and $\mathbf{V}$ is a $p\times n$ matrix with orthonormal columns. Consider a standard regression model, and let $\mathbf{y}$ denotes an n-dimensional response vector, $\mathbf{X}$ denotes an $n\times p$ predictor matrix; $\boldsymbol{\beta}$ denotes a p-dimensional vector of unknown regression parameters, and $\boldsymbol{\varepsilon}$ denotes a random vector with zero mean and one variance. Then we have:

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{UDV}^T\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{H}\boldsymbol{\gamma}+\boldsymbol{\varepsilon} \tag{5}$$

Here $\mathrm{H=UD}$ and $\boldsymbol{\gamma}=\mathbf{V}^{\mathrm{T}}\boldsymbol{\beta}$. So the model descript in (5) can lead to an estimate of $\boldsymbol{\beta}$ by multiplying $\mathbf{V}$ to the least squares estimator of $\boldsymbol{\gamma}$ in equation (5). Then the regression model is applied to the classification problem using the optimal scoring algorithm, listed as follows.

First, let g_{i} denote the tumor class for the ith sample (i=1,…,n). Assuming there are G tumor classes, so that g_{i} takes values {1,…,G}. Then convert $g=[g_1,\ldots,g_{\mathrm{n}}]^{\mathrm{T}}$ into an $n\times G$ matrix Y=[Y_{ij}], where Y_{ij} =1 if the ith sample falls into class j, and 0 otherwise. Let $\theta_{\mathrm{k}}(g)=[\theta_{\mathrm{k}}(g_1),\ldots,\ \theta_{\mathrm{k}}(g_{\mathrm{n}})]^{\mathrm{T}}$ (k=1,…,G) be the $n\times 1$ vector of quantitative scores assigned to g for the kth class.

Step 1: Choose an initial score matrix $\Theta_{G\times J}$ with $J\leq G$-1 satisfying $\Theta^T D_p\Theta=I$, where $D_p=Y^TY/n$. Let $\Theta_0=\mathbf{Y}\Theta$.

Step 2: Fit a multivariate penalized regression model of Θ_0 on A', yielding the fitted values $\hat{\Theta}_0$ and the fitted regression function $\hat{\eta}_0(\mathbf{A})$, and then minimize the function $ASR=\frac{1}{n}\sum_{k=1}^{G}\sum_{i=1}^{n}(\theta_k(g_i)-\mathbf{X}_i^T\beta_k)^2$. Let $\hat{\boldsymbol{\eta}}(\mathbf{X})=S^+\hat{\eta}_0(\mathbf{A})$ be the vector of the fitted regression function on X, where S^+ is the pseudo inverse of S.

Step 3: Obtain the eigenvector matrix Φ of $\Theta_0^T\hat{\Theta}_0$, and hence the optimal scores $\Theta_1=\Theta\Phi$.

Step 4: Let $\eta(\mathbf{X})=\Phi^T\hat{\eta}(\mathbf{X})$.

Step 5: Use the nearest centroid rule to form the classifier, assign a new sample $\mathrm{X_{new}}$ to the class j that minimizes: $\delta(\mathbf{X}_{new},j)=\left\|\mathbf{D}(\eta(\mathbf{X}_{new})-\overline{\eta}^j\right\|^2$ ($\overline{\eta}^j=\frac{\sum_{g_i=j}\eta(\mathbf{X}_i)}{n_j}$ denotes the fitted centroid of the jth class, $D_{kk}=(1/\lambda_k^2(1-\lambda_k^2))^{1/2}$, λ_k is the kth largest eigenvalue calculated in Step 3).

The optimal scoring algorithm is employed to implement the classification by constructing the centroid of the classifier. The algorithm produces a discriminant rule for classifying new samples, and the nearest centroid rule is used to form the classifier.

(b) The ICA+SVM model

For this model, the principle is based on the following analysis.

The original training data sets $\mathrm{X_{tn}}$ and test data sets $\mathrm{X_{tt}}$ are transposed so that they can be applied to evaluate the independent components with the following formulae:

$$U=\mathrm{W_{tn}X_{tn}}=A_{\mathrm{tn}}^{-1}\mathrm{X_{tn}} \tag{6}$$

$$\mathrm{X_{tn}}=A_{\mathrm{tn}}U \tag{7}$$

The rows of A_{tn} contain the coefficients of the linear combination of statistical sources that comprise $\mathrm{X_{tn}}$. Then the representation of the test set $\mathrm{X_{tt}}$ can be calculated as:

$$A_{tt}=X_{tt}U^{-1} \quad (8)$$

And after selecting some special ICs, formulas (1-3) and (6-8) will still be applicable by adjusting A_{tn} as $n\times m$, S as $m\times p$ and A_{tt} as $k\times m$ if there are m ICs selected. Then the ICA prediction mode is constructed based on the selected ICs.

In this model, ICA is directly applied to the classification problem by removing the linear correlations and reducing the high dimensional data to a much lower dimension. After ICA transformation, the data can be classified in a subspace due to the data structure. As SVM has been applied to solve the classification problems successfully in many fields[19, 34], it is also used here for the classification of microarray.

We do not describe the two models in details. The interested readers can refer to the literature [14, 15] for further details. In experiments, it was found that IC selection is necessary in each case in spite of different roles that ICs play in the two models.

2.2 The analysis of ICA

FastICA is applied to transform gene expression datasets. After this processing, the GA is applied to select proper IC subsets. To avoid the problem of convergence to local optima, we use the consensus source based searching algorithm [18] because it may yield stable and robust estimates for the eigenassays. With this method, the independent source estimate is run several times with different random initializations so as to obtain several IC sets. By doing so, the ICs in one IC set are usually quite different from those in other IC sets. Usually, some ICs will appear in different IC sets many times, but some may appear only once. This method only conserves the eigenassays with a frequency larger than a certain threshold, and the appearance frequencies of these eigenassays are used as the *credibility* indices to evaluate the stability of the results. For the i-th IC, the credibility is calculated by

$$\text{credibility}_i=a_i/n \quad (9)$$

where a_i is the frequency that the i-th IC appears in all IC sets, and n is the total number of all the IC sets [15].

However, as pointed out in [18], low credibility doesn't leads to low biological significance, and vice versa. So the frequency of the IC is not a proper criterion for IC selection. And the ICs with higher credibility can't be a guidance of finding the best ICs for classification problem. In other words, it is possible that the relevant ICs related to the classification information would be marked with lower credibility.

2.3 The design of GA

The rise of GA is inspired by mechanisms of evolution in nature. GA has been proved to be successful at tackling the optimization or feature selection problem, so it is a promising solution. The GA applied here is the standard GA, and can be outlined as follows.

Binary coding scheme is applied, and the length of chromosome is equal to the number of ICs. Each gene is valued as 1/0 to represent whether a corresponding IC is/isn't selected. In this way, a chromosome represents a selection mask. The first population is randomly generated. During the decoding process, each chromosome represents a set of selected ICs A'. For the two ICA models, the decoding methods are completely different and can be realized by applying the selected IC subset to the corresponding models.

In the framework of the GA, the selection operator is roulette, which allows individuals with low fitness value to get a chance to enter the next generation. Double point recombination operator is used to exchange a randomly selected part of individuals in pairs. The simple inversion mutation is adopted as the mutation operator, which can randomly select two points in a parent and produces offspring by reversing the genes between the two points. These operators guarantee the diversity among the population.

Leave-one-out cross validation (LOOCV) is applied to evaluate the generalization ability of the classifiers, and the LOOCV accuracy is assigned as the fitness function:

$$C_L=R_L/S_{tn}\times 100 \quad (10)$$

where R_L is the total number of correctly classified samples in the cross validation, and S_{tn} is the total number of samples used for training. As observed in [19], LOOCV easily suffers overfitting, which decreases the generalization ability of the prediction systems. So an early stopping technique, which is formerly used in neural network, is applied to avoid the overfitting in algorithm. To illustrate the process, let IC_t represent the IC set transformed from a training set, and IC_v represent the IC set obtained from a validation set. By applying the mask represented by a chromosome on the IC_t and IC_v, the corresponding IC subset are obtained, which are denoted by IC_t' and IC_v'. The LOOCV accuracy of the training set can be calculated using IC_t', and is assigned as the fitness value of the chromosome. Then IC_t' is applied to train the classifier, which is tested based on IC_v' to evaluate the accuracy on the validation set. This accuracy is not used to change

the corresponding fitness value. Instead, it is used to monitor the change of the LOOCV accuracy on the training set. If the average accuracy of the validation set keeps decreasing greatly in spite that the LOOCV accuracy of the training set keeps increasing, it is obvious that overfitting occurs and the GA should be stopped.

To determine in which generation the GA should be stopped, we define a trend curve to analyze the trend of the average accuracy on the validation set. Let A_c represent the corresponding average classification accuracy on the validation set in the current generation, and our desired goal is to find out the highest average accuracy on the test set. Let $TR(t)$ be the trend of the average prediction accuracy on the validation set in the t-th generation. Based on measuring the changes of A_c in nine generations, we can define $TR(t)$ as:

$$TR(t)=\sum_{t'=t-4}^{t+4}(A_c(t')/9) \quad (11)$$

The trend is estimated after the fifth generation. Then $TR_{max}(t)$ is defined to be the highest trend point obtained up to the t-th generation.

$$TR_{max}(t)=\max_{t'\le t} TR(t') \quad (12)$$

Then the generalization loss is defined to be the relative decrease of the average validation set accuracy over the maximum-so-far, which is similar to the one proposed by Prechelt [20]:

$$GL(t) = 1- TR(t)/TR_{max}(t) \quad (13)$$

GA stops when GL is larger than a threshold. And in our experiments, we set the threshold as 0.05. In this way, if the trend line drops and reaches a point lower than 95% of the highest trend point, it is considered that the overfitting occurs and the GA stops.

After that step, the algorithm goes back to the generation in which the highest value of the trend curve is achieved. Although it is intuitive to use the best individual in this generation, we find that sometimes there are more than one individual leads to the same accuracy rate in the training and validation sets. At the same time, due to the small size of validation sets, the highest accuracy on the validation set may not always guarantee the best performance on the test set. Instead, it has been proved that a multiple classifier system is more robust than an excellent classifier in many fields, and the application of ensemble system in microarray datasets has been proved to be successful [21-23]. So once the highest point of the trend curve is determined, all individuals in the corresponding generation are used to build base classifiers. All base classifiers are combined to construct an ensemble system finally. It should be noted that a robust ensemble system should contain accurate and diverse classifiers. That is, the base classifiers should be of high classification accuracy and avoid making coincident errors. In this way, a sample misclassified by a base classifier will be corrected by others, so the fused outputs are more accurate than that of the best individual classifier. And no gains will be achieved when fusing classifiers producing the same outputs. As the difference among the individuals can guarantee the diversity among the classifiers, the ensemble system built in this way usually achieves good performance. Many feature selection algorithms, such as SFS and SFFS, always select a feature subset to train classifier. As a result, they can't be used to generate an ensemble system.

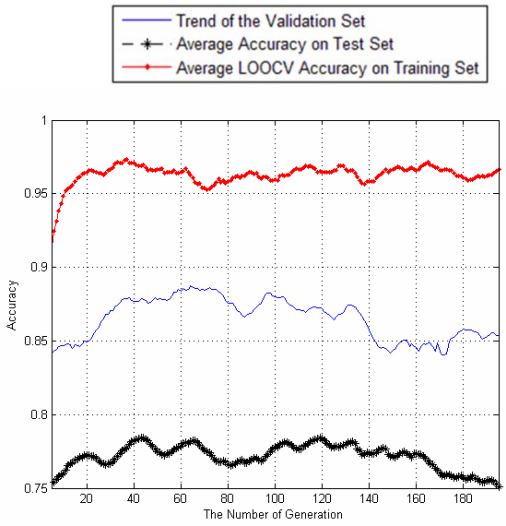
3 Experimental results and discussions

We use three publicly available microarray datasets for comparisons with other research works: the colon cancer dataset [24], the hepatocellular carcinoma dataset [25] and the high-grade glioma dataset [26]. In these datasets, all data samples have already been assigned to the training set or test set. Preprocessing of these datasets is done by setting threshold and log-transforming on the original data, similar to the original publication. Threshold technique is generally achieved by restricting gene expression levels to be larger than 20 and those smaller than 20 will be set to 20. Regarding the log-transformation, the natural logarithm of the expression levels usually is taken. No further preprocessing is applied to the datasets. The details about the datasets are listed in Table 1.

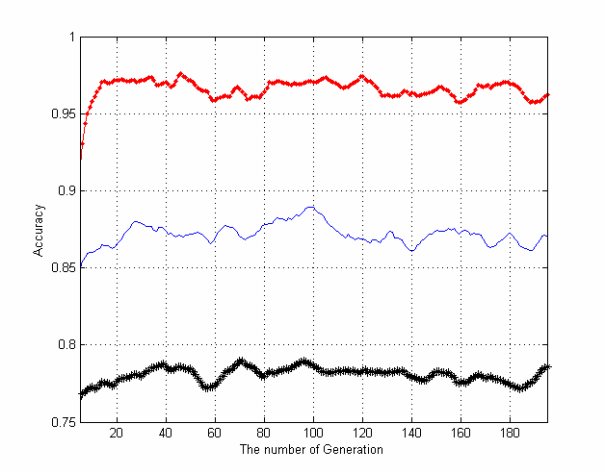
Our GA is based on the GEATbx [27]. When using the GA to classify the tumor data, the original training sets are randomly split to a training subset and a validation subset in proportion of 2:1. The samples are stratified when assigned to the training subset and validation subset, that is, the training subset and validation subset contain the same proportion of samples of each class compared to the original training set and test set. When GA stops, the individuals in the selected generation are all used to build classifiers. The classification accuracies on the training set and test set are estimated based on the original training set and test set. In the GA, there are 50 chromosomes in a generation, so the ensemble system contains 50 base classifiers. The GA stops when it runs 200 generations or the overfitting is detected. The generation gap is 0.9, and the rate for crossover and mutation is set to 0.7 and 1, respectively.

For comparisons, SFFS is also applied to deal with the IC selection problem in experiments. It should be noted that SFFS evaluates the features one by one, and GA always treat a feature subset as a whole. It should be noted that after adding the new feature, the new subset may not necessarily output higher accuracy than the original one. So a pruning process would be carried out to leave out an IC if without the IC the remained IC subset could achieve higher prediction accuracy. After pruning, the performance can be improved. For SFFS, it was pointed out that the original SFFS may end up with a variable subset that is worse than the one found before backtracking [28]. So the framework of the SFFS used in this paper is slightly different from the original. The difference lies in that when the SFFS ends its search process, the algorithm backtracks to the IC subset, which achieves the highest LOOCV prediction accuracy. The SFFS terminates when the number of selected features reaches a certain number, which is determined by the LOOCV. The classification models are built using the training samples, and the classification correct rates are estimated using the test set.

The algorithms are tested with running ICA algorithm with randomized initiations. And after ICA transformation, the number of ICs is always set equal to the number of samples in the training sets so as to simplify our discussion. In each experiment, we run the GA and SFFS algorithms subsequently based on the results of a same ICA transformation to compare their performance. We use the SVM model with RBF kernel function, and the penalty parameter C and the kernel function parameter γ are set to 300 and 5 in all corresponding experiments.



(a)



(b)

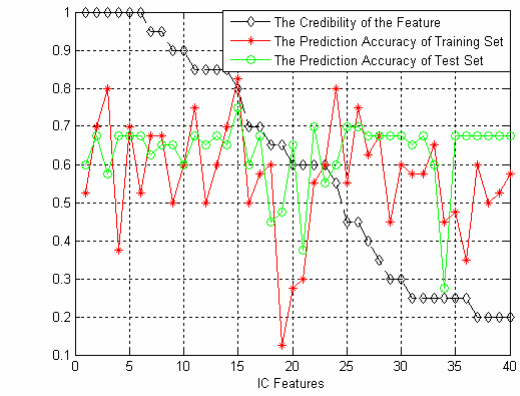
Fig. 1. The examples with or without overfitting case when applying GA for feature selection based on the ICA+SVM model on the colon data sets: (a) Overfiting occurred obviously; (b) Overfiting did not occur obviously.

From Fig.1.(a), we can find that the trend line rose first, then fell slightly after the 64th generation. Although it recovered slightly after some generations, it kept falling down after then, which indicates that overfitting occurred in the run. As the trend line reached the highest value at the 64th generation, the individuals in the 64th generation were all used to build base classifiers. And from Fig.1.(a), it is found that the average classification accuracy on the test set was really the optimal one in this run. So it is obvious that the early stopping can help to find the best generation which leads to the highest average results on the test set before overfitting occurs. In experiments, it is found that the overfitting would be recovered in a slight degree in most cases, which can also be observed in Fig.1.(a). But the best results obtained after the occurrence of overfitting are still worse than those without overfitting. On the other hand, even when overfitting does not occur obviously, the trend curve could at least help to find the near optimal results. As shown in Fig.1.(b), the highest point in the trend line is reached at the 99th generation. And in this generation, the average accuracy on the test set is also close to the highest result.

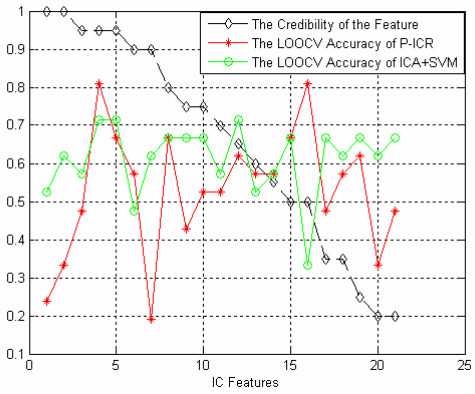
In the experiments, we find that there is no way to completely escape from overfitting. It is impossible for us to foresee whether overfitting would occur or not, so it is necessary and important to apply the early stopping technique. In addition, when the distribution of data in the validation set is quite different from that in the test set, the trend line may fail to locate the optimal results on the test set. However, in our experiments, the trend line works well usually.

According to (11), it should be noted that we assume the best generation could be found between the 5th generation to 195th generation. It is usually the case, and we do not change this assumption in our experiments. It is obvious that we can't obtain good results within five generations. If it is necessary to discuss the results after 195 generations, it can be achieved by adjusting formula (11), or simply running GA with some more generations.

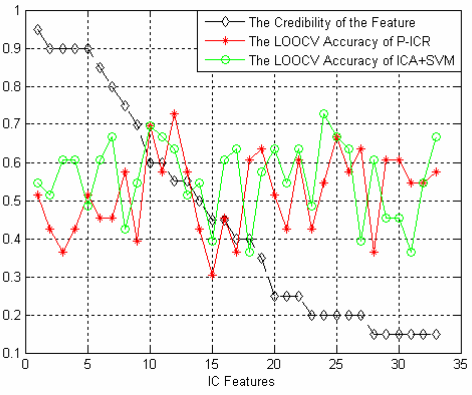
We evaluate the contributions of different ICs by applying LOOCV. The relationship of the LOOCV accuracy and the credibility of the ICs can be illustrated in Fig.2(a-c), where the ICs were sorted by the credibility. From these figures, we can find that the credibility can't indicate the classification accuracy directly. The LOOCV accuracy vibrates greatly in spite of the decrease of the credibility curve. And it should be noted that for the two different ICA based classification models, the LOOCV accuracy of each IC is different, which indicates that different ICs should be selected for different prediction models. So it is impossible to set up a general rule for IC selection based on different prediction models. In addition, the filter selection strategy, which is independent of prediction systems, is not a proper solution for tackling the IC selection problem.



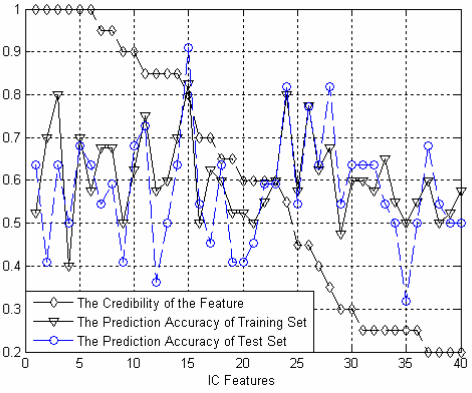
(a)



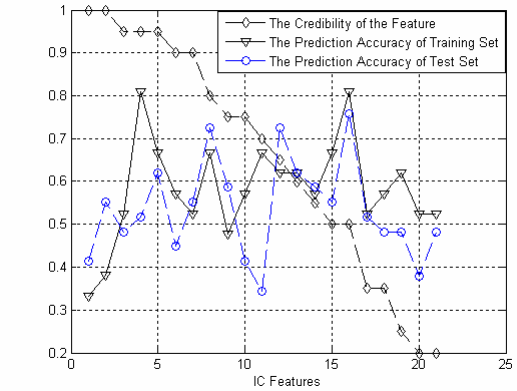
(b)



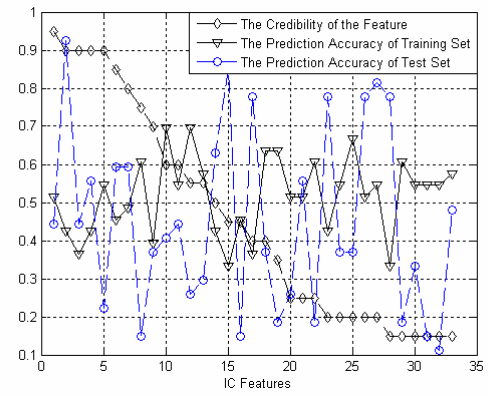
(c)



(d)



(e)



(f)

Fig. 2. The accuracy *vs*. the credibility: LOOCV classification accuracy on the training data set *vs*. the credibility on all data set: (a) the Colon dataset; (b) the Glioma dataset; (c) the Hepatocellular dataset. The prediction ability on the training and the test data set *vs*. the credibility on all data set: (d) the Colon dataset; (e) the Glioma dataset; (f) the Hepatocellular dataset.

We also evaluate the classification ability of each IC by applying a single IC to construct the prediction system. As it is impossible to train SVM with a single IC, the evaluation is processed only based on P-ICR model. The corresponding results are shown in Fig.2.(d-f). Again we find that all results, both training data and test data, don't follow the distribution of the credibility. And the ICs with low credibility sometimes result in high classification accuracy, which is consistent with the finding in [6]. In conclusion, the credibility of the ICs can only indicate that the distribution of ICs is close to nongaussian or not, and does not affect the classification accuracy directly.

For comparison, we list the results using 11 different methods: PCA and kernel PCA with FDA, LS-SVM [19]; the P-PCR [29], and PAM [30] in Table 2. To evaluate the results objectively, we compare both the prediction accuracy on the training and the test sets.

From Table 2, it is obvious that for the two models, without IC feature selection, the results are worse or close to the results obtained with other methods. By selecting proper IC subsets for classification, the accuracies are higher than the originals, and usually higher than others methods. For example, by using the whole set of ICs for classifying the colon dataset, the average test results of P-ICR and ICA+SVM model is only 67.27% and 68.18%. While after selecting proper IC subsets with SFFS and GA, the average accuracy can reach 86.82% and 91.57% for P-ICR model, and 87.91% and 91.52% for ICA+SVM model. The best results of all other 11 methods can only reach 85.54% at most. So it is obvious that the IC selection schemes can improve the performance of the ICA based models, and the GA based scheme can lead to the highest accuracy. And the same conclusion can also be drawn when testing on other datasets, as shown in Table 2. So the IC selection method can produce higher accuracy in both classification and regression models, and usually beats all other methods, we can safely conclude that IC selection is an efficient method for ICA based models, and the GA based scheme works best.

What's more, the results of GA based scheme are superior to those of SFFS based scheme due to the powerful search capability embedded in GA. As the SFFS algorithm can only add or prune an IC at a time, it lacks of the ability of evaluating an IC subset as a whole, which can be achieved by GA. That is the reason that GA has great advantage over the SFFS.

At the same time, it is obvious that the results obtained by the ensemble system are much higher than those of the average accuracies with lower standard deviations. It is because the GA can maintain the diversity among individuals naturally, which guarantees the difference among the base classifiers. In this way, this ensemble scheme is efficient and effective.

In conclusion, the GA based IC selection algorithm is efficient for both models, and it is necessary to further consider the selection of proper ICs when applying ICA based classification models.

4 Conclusion

In this paper, we discussed the IC selection problem for classification of tumors based on microarray gene expression data, and designed a GA based selection scheme to implement the selection on two different ICA based models. LOOCV is applied to evaluate the selected ICs, and an early stopping technique is designed to overcome the overfitting caused by LOOCV. An optimal population is determined by early stopping technique, and then it is used to construct an ensemble system. With this scheme, we find that the GA is effective and efficient in predicting normal and tumor samples from the three human tissues. So it is obvious that IC selection is necessary for better classification results.

In future works, we will further and deep study the ICA model of gene expression data, and try to discover the relationship of different ICs and what the role of the credibility is playing in the classification problem.

Acknowledgement

Supported by the Natural Science Foundation of Fujian Province of China (No.2010J05137), the Fundamental Research Funds for the Central Universities(No.2010121038) and Leading Academic Discipline Program, 211 Project for Xiamen University (the 3rd phase).

References:

1. West, M., *Bayesian Factor Regression Models in the 'Large p, Small n' Paradigm.* Bayesian Statistics, 2003. 7: p. 723-732.
2. Bartlett, M.S., Movellan, J.R. and Sejnowski, T.J., *Face recognition by independent component*

analysis. Ieee Transactions on Neural Networks, 2002. 13(6): p. 1450-1464.
3. Liebermeister, W., *Linear modes of gene expression determined by independent component analysis*. Bioinformatics, 2002. 18: p. 51-60.
4. Lee, S.I. and Batzoglou, S., *Application of independent component analysis to microarrays*. Genome Biol., 2003. 4(R76).
5. Zhang, X.W., Yap, Y.L., Wei, D., Chen, F. and Danchin, A., *Molecular Diagnosis of Human Cancer Type by Gene Expression Profiles and Independent Component Analysis.* European Journal of Human Genetics, 2005. 13(12): p. 1303-1311.
6. Frigyesi, A., Veerla, S., Lindgren, D. and Hoglund, M., *Independent component analysis reveals new and biologically significant structures in microarray data.* BMC Bioinformatics 2006. 7:290.
7. Elashoff, J.D., Elashoff, R.M. and Goldman, G.E., *On the choice of variables in classification problems with dichotomous variables,, , , .* Biometrika, 1967. 54: p. 668-670.
8. Macias-Macias, M., Garcia-Orellana, C.J., Gonzalez-Velasco, H. and Gallardo-Caballero, R., *ICA and GA feature extraction and selection for cloud classification.* Pattern Recognition and Data Mining, Pt 1, Proceedings, 2005. 3686: p. 488-496.
9. Zheng, C.H., Huang, D.S. and Shang, L., *Feature selection in independent component subspace for microarray data classification.* Neurocomputing, 2006. 69(16-18): p. 2407-2410.
10. Peterson, D.A., Knight, J.N., Kirby, M.J., Anderson, C.W. and Thaut, M.H., *Feature selection and blind source separation in an EEG-based brain-computer interface.* Eurasip Journal on Applied Signal Processing, 2005. 2005(19): p. 3128-3140.
11. Ekenel, H.K. and Sankur, B., *Feature selection in the independent component subspace for face recognition.* Pattern Recognition Letters, 2004. 25(12): p. 1377-1388.
12. Huang, Y.P. and Luo, S.W., *Genetic algorithm applied to ICA feature selection.* In: proceedings of international joint conference of neural network, 2003: p. 704-707.
13. Kudo, M. and Sklansky, J., *Comparison of algorithms that select features for pattern classifiers.* Pattern Recognition, 2000. 33(1): p. 25-41.
14. Zheng, C.H., Chen, Y., Li, X.X., Li, Y.X. and Zhu, Y.P., *Tumor classification based on independent component analysis.* International Journal of Pattern Recognition and Artifical Intelligence 2006. 20(2): p. 297-310.
15. Huang, D.S. and Zheng, C.H., *Independent component analysis-based penalized discriminant method for tumor classification using gene expression data.* Bioinformatics, 2006. 22: p. 1855-1862.
16. Hastie, T., Tibshirani, R. and Buja, A., *Flexible discriminant analysis by optimal scoring.* Journal of the American Statistical Association, 1994. 89: p. 1255-1270.
17. Hyvärinen, A., *Fast and robust fixed-point algorithms for independent component analysis.* IEEE Trans. Neural Netw., 1999. 10: p. 626–634.
18. Chiappetta, P., Roubaud, M.C. and Torresani, B., *Blind source separation and the analysis of microarray data.* Journal of Computational Biology, 2004. 11: p. 1090-1109.
19. Pochet, N., De Smet, F., Suykens, J.A.K. and De Moor, B.L.R., *Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction.* Bioinformatics, 2004. 20(17): p. 3185-3195.
20. Prechelt, L., *Automatic early stopping using cross validation: quantifying the criteria.* Neural Networks, 1998. 11(44): p. 761-767.
21. Diaz-Uriarte, R. and Alvarez de Andres, S., *Gene selection and classification of microarray data using random forest.* Bmc Bioinformatics, 2006. 7(3).
22. Liu, K.-H. and Huang, D.-S., *Cancer classification using Rotation Forest.* Computers in Biology and Medicine, 2008. 38(5): p. 601-610.
23. Nanni, L. and Lumini, A., *Ensemblator: An ensemble of classifiers for reliable classification of biological data.* Pattern Recognition Letters, 2007. 28(5): p. 622-630.
24. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J., *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, 1999. 96(12): p. 6745-6750.
25. Iizuka, N., Oka, M., Yamada-Okabe, H., Nishida, M., Maeda, Y., Mori, N., Takao, T., Tamesa, T., Tangoku, A. and Tabuchi, H., *Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection.* The Lancet, 2003. 361: p. 923–929.
26. Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Cairncross, J.G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M.E., Batchelor, T.T., Black, P.M., von Deimling, A., Pomeroy, S.L., Golub, T.R. and Louis, D.N., *Gene expression-based classification of malignant gliomas correlates*

better with survival than histological classification. Cancer Research, 2003. 63(7): p. 1602-1607.
27.Pohlheim, H., *GEATbx - Genetic and Evolutionary Algorithm Toolbox for use with Matlab.* 1994-2006.
28.Somol, P., Pudil, P., Novovicova, J. and Paclik, P., *Adaptive floating search methods in feature selection.* Pattern Recognition Letters, 1999. 20(11-13): p. 1157-1163.
29.Ghosh, D., *Penalized discriminant methods for the classification of tumors from microarray experiments.* Biometrics, 2003. 59: p. 992-1000.
30.Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G., *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* PNAS, 2002. 99: p. 6567-6572.
31. Liu, K.H., et al., *Microarray data classification based on ensemble independent component selection*, Computers in Biology and Medicine. Vol. 39, No.11, pp. 953 – 960, 2009.
32. Liu, K.H., et al., *A New Approach to Improving ICA-Based Models for the Classification of Microarray Data*, International Symposium on Neural Networks (ISNN), Hubei, Wuhan, May 26-29, 2009, LNCS 5553, pp. 983–992
33. Liu, K.H., et al., *Ensemble component selection for improving ICA based microarray data prediction models*, Pattern Recognition, Vol. 42, No. 7, pp. 1274–1283, 2009.
34. Shao, S.Y., et al, *Automatic EEG Artifact Removal: A weighted Support Vector Machine Approach with Error Correction*, IEEE Trans. on Biomedical Engioneering, Vol. 56, No.2, pp. 336-344, 2009.

Table 1. The summary of the datasets

Datasets	Original Training Set		Original Test Set		the number of genes	microarray technology
	class 1	class 2	class 1	class 2		
Colon cancer data	14	26	8	14	2000	Oligonucleotide
Hepatocellular carcinoma data	12	21	8	19	7129	Oligonucleotide
High-grade glioma data	21	14	14	15	12625	Oligonucleotide

Table 2. The results of the numerical experiments on three datasets. Here, results listed in the row 'average' in the 15th and 19th methods are the average accuracy rates of base classifiers in the corresponding ensemble system.

Experiments		Colon data		Hepatocellular data		Glioma data	
No.	Methods	Training set	Test set	Training set	Test set	Training set	Test set
1	LS-SVM linear kernel	99.64±0.87	82.03±7.49	73.88±16.21	68.43±4.52	90.02±14.16	61.25±11.75
2	LS-SVM RBF kernel	98.33±2.36	81.39±9.19	87.16±16.73	68.61±6.32	98.41±7.10	69.95±8.59
3	LS-SVM linear kernel (no regularization)	49.40±8.93	51.73±12.19	53.82±5.68	49.56±12.60	50.79±12.75	48.93±10.88
4	PCA + FDA (unsupervised PC selection)	90.95±5.32	80.30±9.65	89.61±9.92	68.25±7.37	92.29±7.12	68.72±7.24
5	PCA + FDA (supervised PC selection)	95.24±5.56	76.84±7.41	90.33±11.52	66.67±9.96	92.97±10.14	65.52±11.01
6	kPCA lin + FDA (unsupervised PC selection)	90.95±5.32	80.30±9.65	89.61±9.92	68.25±7.37	92.52±6.98	68.31±6.78
7	kPCA lin + FDA (supervised PC selection)	95.24±5.56	76.84±7.41	90.33±11.52	66.67±9.96	95.24±8.57	67.32±11.04
8	kPCA RBF + FDA (unsupervised PC selection)	87.86±11.24	75.11±15.02	87.45±12.27	61.20±12.91	94.78±9.05	64.20±11.19
9	kPCA RBF + FDA (supervised PC selection)	100.00±0.00	64.07±1.94	100.00±0.00	69.49±3.94	96.15±7.29	58.13±12.24
10	P-PCR	91.25±2.02	85.54±4.45	92.69±9.97	57.41±8.80	93.33±8.16	70.35±8.19
11	PAM	91.50±4.29	83.63±5.82	89.35±3.64	59.26±9.22	98.57±2.17	67.24±6.58
12	P-ICR	78.00±10.59	67.27±8.24	57.27±11.65	56.67±9.29	92.38±11.04	68.70±5.01
13	P-ICR(SFFS)	95.75±3.81	86.82±6.59	85.67±8.54	72.76±8.93	96.19±4.37	76.76±3.43
14	P-ICR(GA)	97.47±1.06	91.57±4.25	90.53±4.53	75.42±4.35	96.28±3.83	78.37±4.35
15	average	95.22±1.25	80.23±5.23	85.79±4.82	71.13±5.33	94.32±5.63	70.12±5.45
16	ICA+SVM	100.00±0.00	68.18±10.92	100.00±0.00	64.08±7.21	100.00±0.00	68.96±7.62
17	ICA+SVM(SFFS)	100.00±0.00	87.91±5.44	100.00±0.00	72.96±3.05	100.00±0.00	77.59±3.73
18	ICA+SVM(GA)	100.00±0.00	91.52±1.32	100.00±0.00	81.23±3.33	100.00±0.00	79.13±4.62
19	average	100.00±0.00	79.89±4.86	100.00±0.00	74.84±6.52	100.00±0.00	74.34±6.20