Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

# Parameters selection in gene selection using Gaussian kernel support vector machines by genetic algorithm*

MAO Yong (毛 勇)[†1], ZHOU Xiao-bo (周晓波)[†2], PI Dao-ying (皮道映)[†‡1],
SUN You-xian (孙优贤)[1], WONG Stephen T.C.[2]

([1]*National Laboratory of Industrial Control Technology, Institute of Modern Control Engineering, Zhejiang University, Hangzhou 310027, China*)
([2]*Harvard Center for Neurodegeneration and Repair, Harvard Medical School and Brigham and Women's Hospital, Harvard Medical School, Harvard University, Boston, MA 02115, USA*)
[†]E-mail: ymao@iipc.zju.edu.cn; zhou@crystal.harvard.edu; dypi@iipc.zju.edu.cn
Received Dec. 8, 2004; revision accepted Mar. 11, 2005

**Abstract:** In microarray-based cancer classification, gene selection is an important issue owing to the large number of variables and small number of samples as well as its non-linearity. It is difficult to get satisfying results by using conventional linear statistical methods. Recursive feature elimination based on support vector machine (SVM RFE) is an effective algorithm for gene selection and cancer classification, which are integrated into a consistent framework. In this paper, we propose a new method to select parameters of the aforementioned algorithm implemented with Gaussian kernel SVMs as better alternatives to the common practice of selecting the apparently best parameters by using a genetic algorithm to search for a couple of optimal parameter. Fast implementation issues for this method are also discussed for pragmatic reasons. The proposed method was tested on two representative hereditary breast cancer and acute leukaemia datasets. The experimental results indicate that the proposed method performs well in selecting genes and achieves high classification accuracies with these genes.

**Key words:** Gene selection, Support vector machine (SVM), Recursive feature elimination (RFE), Genetic algorithm (GA), Parameter selection
**doi:**10.1631/jzus.2005.B0961 **Document code:** A **CLC number:** Q789; R73

## INTRODUCTION

Recent techniques based on oligonucleotide or cDNA microarrays allow the expression level of thousands of genes to be monitored in parallel (Golub *et al*., 1999). A critically important factor for cancer diagnosis and treatment is the reliable prediction of tumor progression. A remarkable advance for molecular biology and for cancer research is cDNA microarray technology. cDNA microarray datasets have a high dimensionality corresponding to the large number of genes monitored, and there are often comparatively few samples. In this paper, we address the problem in predicting cancer by using a small subset of important genes from a wide collection of gene expression data.

Since Golub *et al*.(1999) proposed a weighted voting scheme for molecular classification of acute leukemia, many existing machine learning methods have been applied to gene classification problems (Kim *et al*., 2000; Tabus and Astola, 2001; Zhou *et al*., 2003a; Alizadeh *et al*., 2000; Mao *et al*., 2004). Given the thousands of genes but the small amount of data samples, ranking genes according to their importance in contributing to classifiers' predictive accuracy is

‡ Corresponding author
* Project supported by the National Basic Research Program (973) of China (No. 2002CB312200) and the Center for Bioinformatics Program Grant of Harvard Center of Neurodegeneration and Repair, Harvard Medical School, Harvard University, Boston, USA

crucial (Zhou *et al*., 2003b; 2003c; 2004a).

Support vector machines (SVMs) are considered a good classification method for gene-expression data and are embedded with feature selection procedures (Cristianini and Shawe-Taylor, 2000; Guyon *et al*., 2002; Weston *et al*., 2001; Zhang and Wong, 2001; Furlanello *et al*., 2003). Recursive feature elimination based on SVM (SVM RFE) discussed in (Guyon *et al*., 2002; Zhang and Wong, 2001; Furlanello *et al*., 2003) is considered a good method in this field. As mentioned in Shashua and Wolf (2004), if feature selection is embedded into a higher dimensional space using a right kernel function, it may be possible to emphasize certain aspects of the data while de-emphasizing the others so that a more reasonable feature selection may be done. We attempt to use recursive feature elimination based on SVM with Gaussian kernel to select more important genes. In the implementation of the aforementioned algorithm, the selection of the model parameters, namely, the width of SVM kernel and penalty parameter, are not always mentioned: empirical parameters are used frequently as in Guyon *et al*.(2002).

This paper proposes a strategy of using a genetic algorithm to search for a couple of parameters which could optimize the results of Gaussian kernel SVM RFE. Since this method has high computational complexity, we also discuss some numerical techniques to speed up the computation for pragmatic implementation. Furthermore, a gene pre-selection procedure is adopted to reduce the huge number of genes being considered for selection. We demonstrate the method on two realistic public domain gene expression datasets, i.e., AML/ALL (Acute Myeloblastic Leukemia/Acute Lymphocytic Leukemia) dataset (Guyon *et al*., 2002) and hereditary breast cancer dataset (Hedenfalk *et al*., 2001). The experimental results showed that the proposed methods can effectively find important genes consistent with the biological considerations, while achieving high classification accuracy.

## PROBLEM FORMULATION

Assume there exist two classes of cancers. Let $\boldsymbol{Y}=[y_1, \ldots, y_m]^{\mathrm{T}}$ denote the class labels of $m$ samples, where $y_i=k$ indicates the sample $i$ being cancer $k$, where $k$=1, 2 denotes two different kinds of cancer (in our experiments, we use $y_i=-1$ to indicate the sample $i$ being cancer 1, and $y_i=1$ to indicate the sample $i$ being cancer 2). Let $x_{ij}$ be the measurement of the expression level of the $j$th gene for the $i$th sample, where $j$=1, 2, …, $n$, $\boldsymbol{X}=(x_{ij})_{m,n}$ denotes the expression levels of all genes, i.e.,

$$\boldsymbol{X}=\begin{bmatrix} \text{Gene 1} & \text{Gene 2} & \cdots & \text{Gene } n \\ x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{1}$$

Here we assume $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are the $m$ samples, where $\boldsymbol{x}_i=[x_{i1}, x_{i2}, \ldots, x_{in}]$.

In our method, every sample is partitioned by an optimal hyper-plane, with training data being maximally distant from the hyper-plane itself. The lowest classification error rate will be achieved when this hyper-plane is used to classify the current training set. This hyper-plane can be modelled as

$$f(\boldsymbol{x})=\sum_{i=1}^{m} y_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x})+b \tag{2}$$

where $\alpha_i$ is the weight of the $\boldsymbol{x}_i$; $b$ is a bias term and $K(\boldsymbol{x}_i, \boldsymbol{x})=\exp\left(-\dfrac{\|\boldsymbol{x}_i-\boldsymbol{x}\|^2}{2\sigma^2}\right)$ is Gaussian radius basis function, $\sigma$ is positive Gaussian kernel width. SVM, a machine learning algorithm originally introduced by Vapnik (2000) was used to determine these optimal hyper-planes. It solves a convex quadratic programming problem to get the optimal values of $\alpha_i$ and $b$ for details of the SVM learning algorithm (Vapnik, 2000).

In Eq.(1), since $n$ is too large, it induces many estimation error problems (Zhou *et al*., 2003a; Guyon *et al*., 2002). So the dimensionality reduction will be done on input data, $\boldsymbol{X}$, from the strongest genes selected. Using function $\tilde{\boldsymbol{x}}_i^{\mathrm{T}}=I(\boldsymbol{\beta}\boldsymbol{x}_i^{\mathrm{T}})$ to represent this procedure, where $\boldsymbol{\beta}$ is a $n\times n$ matrix, in which only diagonal elements may be equal to 1 or 0, and all other elements are equal to zero, genes corresponding to the non-zero diagonal elements are important. The matrix $\boldsymbol{\beta}$ is obtained by specific gene selection

methods given in the next two sections. The function $I(\cdot)$ means to select all non-zero elements in the input vector to construct a new vector, e.g., $I([1\ 0\ 2]^{\mathrm{T}})=[1\ 2]^{\mathrm{T}}$. So Eq.(2) is rewritten as

$$\begin{aligned} &f(\boldsymbol{x})=\sum_{i=1}^{m} y_i\alpha_i K(\tilde{\boldsymbol{x}}_i,\tilde{\boldsymbol{x}})+b;\\ &\tilde{\boldsymbol{x}}_i^{\mathrm{T}}=I(\boldsymbol{\beta}\boldsymbol{x}_i^{\mathrm{T}}),\ \boldsymbol{x}^{\mathrm{T}}=I(\boldsymbol{\beta}\boldsymbol{x}^{\mathrm{T}}). \end{aligned} \tag{3}$$

SVM can be embedded with feature selection procedures. To describe our method freely, it is necessary to briefly describe the method in (Guyon *et al*., 2002). When a Gaussian kernel SVM is trained, the two parameters $C$ and $\sigma^2$ should be pre-fixed: $C$ is the penalty parameter used in the SVM algorithm. In training the SVM with the pre-fixed parameters $C$ and $\sigma^2$, the cost function is defined as

$$J(\boldsymbol{\alpha})=(1/2)\,\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{\alpha}-\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} \tag{4}$$

where $\boldsymbol{H}=(H_{ij})_{i,j=1,\ldots m}$; $H_{ij}=y_iy_jK(x_i,x_j)$, $\boldsymbol{\alpha}=(\alpha_i)_{i=1,\ldots m}$, $0\le\alpha_i\le C$. The importance of a gene for the SVM can be defined in terms of its contribution to this cost function, which is computed as $\Delta J(i)=(1/2)(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{\alpha}-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{H}(-i)\boldsymbol{\alpha})$, $\boldsymbol{H}(-i)$ is $\boldsymbol{H}$ with $i$th gene removed. After the puny gene with smallest $\Delta J$ is eliminated, the new SVM will be retrained using $\hat{\boldsymbol{X}}$ which is defined as $\boldsymbol{X}$ with the puny gene removed. This process is then repeated until the most important gene is obtained. This procedure is called recursive feature elimination by SVM (SVM RFE). Using a gene subset on the top of the ranked list, a Gaussian kernel SVM classifier with proper model parameters ($C$, $\sigma^2$) will be constructed.

The model parameters ($C$, $\sigma^2$) are used in each elimination of SVM RFE. The results of gene selection as well as the construction of classifiers are a direct sequence of the model selected. In what follows, a strategy of selecting model parameters by genetic algorithm is proposed.

## MODEL SELECTION USING GENETIC ALGORITHM IN GAUSSIAN KERNEL SVM RFE

Empirical parameters are used frequently in SVM RFE, because of two aspects. First, computing complexity is too high to optimize the model parameters (Guyon *et al*., 2002), and second, selecting a practically optimal target is difficult. As to the computing complexity, the entropy-based SVM RFE is discussed in (Zhang and Wong, 2001; Furlanello *et al*., 2003), a chunk of features can be eliminated this time according to an entropy-based criterion, and this greatly improves the calculating rate of the original method in (Guyon *et al*., 2002). As to the optimal goal, Furlanello *et al*.(2003) said that a classifier's predictive accuracy is achieved by either double cross-validation or the bootstrap re-sampling process can be used as evaluation criteria in selecting the model used in SVM RFE. But few published literature on this subject exists. Our method is also partly motivated by the report of the discovery of very few genes (2~25 genes) with classifier performance of negligible or zero error rates (Li and Yang, 2002).

In order to use genetic algorithm to optimize the model used in SVM RFE, certain problems must be solved first. Note that the SVM must be retrained after every elimination operation, because the importance of a feature with medium or low importance may be promoted by removing a correlated feature. Thus, the computational cost of RFE is a function of the number of variables. Golub *et al*.(1999) used a Matlab implementation of the linear kernel SVM RFE on a Pentium processor which returns a gene ranking in about 15 min for the entire Colon dataset (2000 genes, 62 patients) and 3 h on the Leukemia dataset (7129 genes, 72 patients) after performing a preprocessing step to reach a fixed number of genes. This means that if this process repeats hundreds of times, it would require more than half a month of processing time. The computational cost thus is the first problem we should address. The second problem in SVM RFE is that there may be some features whose $\Delta J$ are very close. If only $\Delta J$ is used to eliminate features, the result of gene ranking will not be unique (Guyon *et al*., 2002; Zhang and Wong, 2001; Furlanello *et al*., 2003), and some important genes will be lost.

To accelerate the gene selection process and solve the first problem mentioned above, we use $F$-test as an assistant method. The ratio defined by Eq.(5) is used to pre-select genes.

$$R(j)\overset{\Delta}{=}\frac{(\sum_{i=1}^{m}\sum_{k=1}^{2}1_{(y_i=k)}(\bar{\boldsymbol{x}}_{kj}-\bar{\boldsymbol{x}}_j)^2)}{(\sum_{i=1}^{m}\sum_{k=1}^{2}1_{(y_i=k)}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_{kj})^2)},\ 1\le j\le n \tag{5}$$

Here $\bar{x}_j$ denotes the average expression level of gene $j$ across all samples; $\bar{x}_{kj}$ denotes the average expression level of gene $j$ across the samples belonging to class $k$; and the indicator function $1_\Omega$ is equal to one if the event $\Omega$ is true and zero otherwise. Genes with bigger $R(j)$ are selected. A number of genes (below 200) are pre-selected. Beyond this step, removal of a trunk of $\lambda$ genes is repeated until the number of remaining genes goes below a fixed number $\varsigma$. Then genes will be eliminated one by one. If genes are eliminated one by one after gene pre-selection, by a proper selection of $\lambda$ and $\varsigma$, there will be no remarkable distinction between these two ranked lists, but the ranking time will decrease greatly. Aiming at the second problem, if some genes that may be eliminated have similar $\Delta J$, the gene whose $F$-test value is smaller will be eliminated, which means this particular gene expressed remarkably less by $F$-test estimation. By these measures, a ranking process on Leukemia dataset will take only several minutes to yield a unique ranked list. In our experiments, $\lambda$ was set at 2 and $\varsigma$ at 80.

When $C$ and $\sigma^2$ are used for Gaussian kernel SVM RFE, a ranked list of genes will be obtained. To acquire a set of more meaningful genes, we use a Gaussian kernel classifier to evaluate the ranked list. For the dataset with the remaining genes, the kernel width of the SVM classifier is reconstructed for better evaluation this low-dimensional dataset. We can evaluate the capability of the ranked genes from two different perspectives. First, if this group of genes is ranked appropriately, the most important genes should be in the first several positions. Li and Yang (2002) concluded that it is enough to construct a good classifier with several genes. Similar results were also achieved in (Weston *et al*., 2001; Chapelle *et al*., 2002), in which only five genes were used, and 1~2 errors occurred on the test datasets. If the pre-defined number of genes is used, the evaluating classifier should have good classification performance on the training datasets. To guarantee adequate generalization performance, the evaluating classifier should have a relatively small estimation of the upper bound of the leave-one-out error rate. However a drawback of this evaluation method is that there is no direct relationship among model parameters ($C$, $\sigma^2$), selected genes and performance of the evaluating classifier. To solve this problem, a genetic algorithm is deployed.

Genetic algorithm (GA) is a global stochastic optimization algorithm based on the mechanism of natural selection and natural genetics (Miettinen *et al*., 1999) designed to efficiently search large, non-linear, discrete and poorly understood search space where expert knowledge is scarce or difficult to model while traditional optimization techniques fail. In the remainder of this section, we will discuss in detail how to combine them to select genes in an optimal way. Note that here the optimal solution means that the selection of $C$ and $\sigma^2$ is an optimization problem with constraint, and real-coded scheme of variables is used for higher numerical accuracy than the binary-coded scheme.

**Initialize population**

To accelerate the convergence rate of GA and keep the final solution in a proper sphere, a certain technique should be used here. $C$ is a positive constant (penalty parameter) that can be close to infinity, although $C$=1000 is enough for many operations. An evenly distributed initialization is appropriate for this parameter. Only the magnitude level of $\sigma^2$ is important. If even initialization is used for this parameter, many generations will have to pass before an optimized solution is reached. To initialize it better, a rough estimation of $\sigma^2$ is necessary. We define

$$d_{\text{mean}}^2 = \frac{\sum_{i,j=1;i\neq j}^{m} 1_{(y_i \neq y_j)} \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|^2}{\sum_{i,j=1;i\neq j}^{m} 1_{(y_i \neq y_j)}} \tag{6}$$

where $\boldsymbol{x}_i$ means sample $i$ in current training dataset. $d_{\text{mean}}^2$ is considered as a benchmark for initializing $\sigma^2$. In our experiments, $d_{\text{mean}}^2$ is always near the precise estimation of $\sigma^2$. If population size $M$ is an even number, $2^{-M/2}d_{\text{mean}}^2$, $2^{-(M+2)/2}d_{\text{mean}}^2$, …, $2^{(M-4)/2}d_{\text{mean}}^2$, $2^{(M-2)/2}d_{\text{mean}}^2$ are used as candidates for $\sigma^2$; if $M$ is an odd number, $2^{-(M+1)/2}d_{\text{mean}}^2$, $2^{-(M+3)/2}d_{\text{mean}}^2$, …, $2^{(M-3)/2}d_{\text{mean}}^2$, $2^{(M-1)/2}d_{\text{mean}}^2$ are used as candidates for $\sigma^2$.

**Fitness function**

A ranked list of genes can be obtained by Gaus-

sian kernel SVM RFE from a model $(C, \sigma^2)$ is chosen from the chromosome space. The selected top $\ell$ genes can be used to construct a Gaussian kernel SVM classifier with new kernel width parameter. The new kernel width parameter is also estimated by Eq.(6). We will select the model which yields a better result based on the performance evaluation of these classifiers, and genes ranked with this model are best ordered.

Denote the estimation of the upper bound of generalization error rate of SVM classifier to be $u$, the number of support vectors in the classifier to be $\kappa$ and the rejection rate of the leave-one-out test to be $v$. We next use these three parameters to evaluate the trained SVM. $u$ and $\kappa$ are used to guarantee that the classifier has good generalization performance on test datasets. $v$ is used to guarantee the decision made by this classifier is important enough and believable, especially when the whole training dataset is in an un-separable status. A SVM classifier with smaller $u$, $\kappa$ and $v$ under a given rejection threshold $\eta$ is considered better. $u$ is defined in (Vapnik, 2000) as follows

$$u=\frac{1}{l}\frac{R^2}{\gamma^2} \tag{7}$$

where $R$ is the radius of the smallest sphere enclosing the training points in a high dimensional feature space, $l$ is the size of the training set and $\gamma^2$ is the square of the classifier's margin, which is calculated as $\gamma^2=1/(2\sum_{i=1}^{l}\alpha_i-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{\alpha})$, where $\boldsymbol{H}=(H_{ij})_{i,j=1,\ldots m}$; $H_{ij}=y_iy_jK(x_i,x_j)$, $\boldsymbol{\alpha}=(\alpha_i)_{i=1,\ldots m}$, $0\leq\alpha_i\leq C$. And $R$ is found by solving another quadratic optimization problem (Vapnik, 2000):

$$R^2=\max_{\beta}\left\{\sum_{i=1}^{l}\beta_iK(\boldsymbol{x}_i-\boldsymbol{x}_j)-\sum_{i,j=1}^{l}\beta_i\beta_jK(\boldsymbol{x}_i-\boldsymbol{x}_j)\right\}$$

under constraints $\sum_{i=1}^{l}\beta_i=1$ and $\forall_i\beta_i\geq 0$.

$v$ is a direct result from the leave-one-out test. Given a fixed threshold $\eta$, the total number of samples with a wrong decision or a decision whose absolute value is below $\eta$ is assigned to $v$, which is used as a substitute for extreme margin and medium margin methods mentioned in (Guyon *et al*., 2002). By means of $v$, $u$ and $\kappa$ from the reconstructed SVM classifier, the fitness function of GA can be defined as $fit=(\kappa\cdot(u+\varepsilon))^{-1}\cdot e^{-v}$, where $\varepsilon$ was set at 0.01 in our experiments.

**Genetic operator**

The three basic operators of GA are: selection, crossover and mutation. Ranking method is used here as the selection operator. The probability of the $i$th individual being selected is defined by $p_i=q(1-q)^{r-1}/(1-(1-q)^{\tau})$ (Houck *et al*., 1995), where $q$ is the parameter to control the proportion of the individuals selected; $r$ is the rank of the individuals (the rank is sorted by individual fitness values, where 1 corresponds to the individual with the best fitness), and $\tau$ is the population size. A transformation of arithmetic crossover is used in succession, which is described as $S=r\overline{P}_1+(1-r)\overline{P}_2$, where $r$ is a uniform number between (0,1), and $\overline{P}_1$ and $\overline{P}_2$ are two parents. $S$ and the parent whose fitness is better are the individuals come out in this operator. Using this method, more individuals with better fitness value will be generated than using single arithmetic crossover and more new individuals will be generated than using the heuristic crossover in (Houck *et al*., 1995). For the mutation operator, we use multi-non-uniform mutation to maintain the status of best individuals and the diversity of the whole population, which is also used in (Srinivas and Patnaik, 1994). The non-uniform operator of any of the dimensional variables $P$ in the parent vector can be defined as $B=\begin{cases}P+(b_i-P)g(G) & if\ r_1<0.5\\ P-(P+a_i)g(G) & if\ r_1\geq 0.5\end{cases}$, where $g(G)=(r_2(1-G/G_{\max}))^b$, $r_1$, $r_2$ are uniform numbers between (0, 1), $a_i$, $b_i$ are the upper and lower bounds of this variable, $G$ is the current generation, and $G_{\max}$ is the maximum number of generations. $b$ was 5 in our experiments.

To avoid prematurity of GA (Srinivas and Patnaik, 1994), an adaptive search probability strategy was adopted to improve the quality of genetic optimization for the model selection. Probability for crossover is defined as in (Lee *et al*., 2003),

$$p_c=\begin{cases}\dfrac{k_1(l_{\max}-l')}{l_{\max}-\overline{l}} & l'\geq\overline{l}\\ k_2 & l'<\overline{l}\end{cases},$$

where $l_{\max}$ is the maximal fitness value in the current generation, $l'$ is the bigger fitness value of the two parents. $\overline{l}$ is the mean of the fitness value for the current generation. $k_1$, $k_2$ were 0.8 in our experiments. Probability for mutation is:

$$p_m=\begin{cases}\dfrac{k_3(l_{\max}-l_1)}{l_{\max}-\overline{l}} & l_1\geq\overline{l}\\ k_4 & l_1<\overline{l}\end{cases}$$

where $l_1$ is the fitness value of the parent, $k_3$, $k_4$ were 0.2 in our experiments. Fig.1 summarizes the flow chart of the whole process.

## EXPERIMENTAL RESULTS

Two datasets from hereditary breast cancer and acute leukemia are considered in this work to evaluate the two proposed algorithms.

**Breast cancer dataset**

In our first experiment, we focused on hereditary breast cancer data, which can be downloaded from the web page for the original paper (Hedenfalk *et al*., 2001). In (Hedenfalk *et al*., 2001), cDNA microarrays were used in conjunction with classification algorithms to show the feasibility of using differences in global gene expression profiles to separate BRCA1 and BRCA2 mutation-positive breast cancers. Twenty-two breast tumor samples from 21 patients were examined: 7 BRCA1, 8 BRCA2, and 7 sporadic. There were 3226 genes for each tumor sample. We used our methods to classify BRCA1 versus the others (BRCA2 and sporadic). The ratio data was truncated from below at 0.1 and above at 20. A logarithm operation was performed on the ratio data.

First, 200 important genes were pre-selected by *F*-test. The population size was initialized at 144, in which $C$ was between 0.001 and 1000. About the parameters used in GA, $\ell$ was set at 2, $\eta$ at 0.2, $q$ at 0.05, $k_1$, $k_2$ were 0.8, and, $k_3$, $k_4$ were 0.2. If the fitness

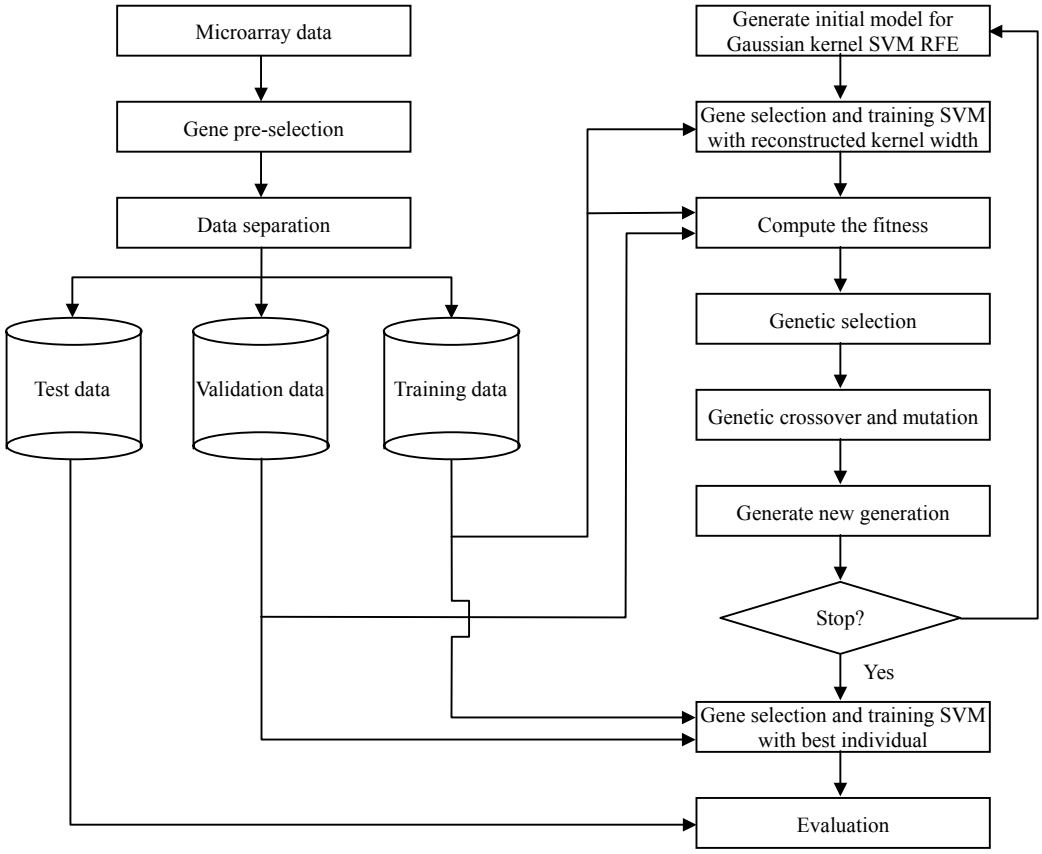


**Fig.1 Flow chart of model selection for gene selection using Gaussian kernel SVM by GA**

value of the best population did not increase in four consecutive generations or the maximal time of generations 20 was reached, GA was terminated. The optimized process is shown in Fig.2 showing that the GA converged after seven iterations, and that the average fitness value was close to the best fitness value. Penalty parameter $C$ fluctuated in a big region between 200 and 900 in the first several generations, but in the final four generations, $C$ changed from 800 to 900, and the kernel width changed slightly.

The top 20 genes selected using the best model produced are listed in Table 1. Gene 336 (TOB1) is also considered an important gene in (Lee *et al*., 2003; Kim *et al*., 2002; Zhou *et al*., 2005; Mao *et al*., 2004).

In order to evaluate the selected genes comprehensively, linear SVM and Gaussian kernel SVM were used as classifiers based on the top 1 to the top 32 genes; these results are listed in Tables 2~3. The parameters used in Tables 2~7 are defined as in (Guyon *et al*., 2002): $V_{suc}$ is the number of samples classified correctly in leave-one-out test at zero rejection, which is used for the common leave-one-out error rate test as well as for the leave-one-out error rate test mentioned in our paper; $V_{acc}$ is maximum number of samples accepted in leave-one-out test to obtain zero error, the rejection threshold lies on the biggest one of the absolute value of false soft-decision; $V_{ext}$ is the difference between the smallest output of the positive class samples and the largest output of the negative class samples (rescaled by the largest difference between outputs); $V_{med}$ is the difference between the median output of the positive class samples and the median output of the negative class samples (rescaled by the largest difference between outputs); $V_{suc}$, $V_{acc}$, $V_{ext}$ and $V_{med}$ were used on training dataset; $T_{suc}$, $T_{acc}$, $T_{ext}$ and $T_{med}$ were evaluating parameters with similar meaning as that used in the test dataset.

In Table 2, linear SVM combined with the top 4 genes, top 8 genes and top 16 genes achieved 0 leave-one-out errors; when the top 2 genes were used, and 1 leave-one-out error was found. Gaussian kernel SVM is used in Table 3. Fig.3 describes training dataset samples obtained by using the top 2 genes, when 0 leave-one-out errors were found. The soft decision results of the leave-one-out test are depicted in Fig.4. It is noteworthy that the soft values of all samples except samples 2 and 4 were very close to the true decision values (−1 and 1). In addition, byusing the top 4 genes, top 8 genes and top 16 genes respectively, 0 leave-one-out error was found.

In (Zhou *et al*., 2004b; 2005), the authors proposed a Bayesian approach combined with several information criteria. Those methods also achieved zero error on this dataset using the top 5 and top 10 genes.

**Acute leukemia dataset**

We have also applied the proposed methods on the leukemia data of (Guyon *et al*., 2002), which are available at (http://www-genome.wi.mit.edu/cgi-bin/cancer/publications/pub). The microarray data contains 7129 human genes, sampled from 72 cases of

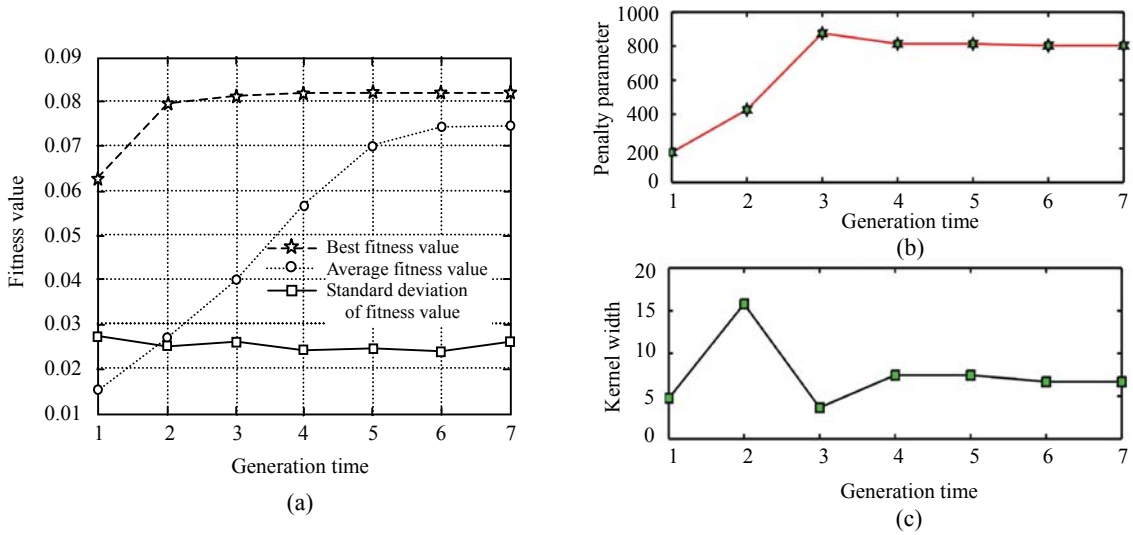


**Fig.2 Model optimization process for gene selection in breast cancer dataset (a) Fitness value; (b) Penalty parameter; (c) Kernel width**

**Table 1 Top 20 genes ranked with optimized Gaussian kernel SVM RFE by GA on breast cancer dataset**

| No. | Index No. | Clone ID. | Gene description |
|---|---|---|---|
| 1 | 2423 | 26082 | Very low lipoprotein receptor |
| 2 | 336 | 823940 | Transducer of ERBB2, 1 (TOB1) |
| 3 | 1999 | 247818 | ESTs |
| 4 | 1620 | 137638 | ESTs |
| 5 | 1277 | 73531 | Nitrogen fixation cluster-like |
| 6 | 1065 | 843076 | Signal transducing adaptor molecule (SH3 domain and ITAM motif) 1 |
| 7 | 498 | 667598 | PC4 and SFRS1 interacting protein 1 |
| 8 | 1008 | 897781 | Keratin 8 |
| 9 | 1288 | 564803 | Forkhead (drosophila)-like 16 |
| 10 | 585 | 293104 | Phytanoyl-CoA hydroxylase (refsum disease) |
| 11 | 2734 | 46019 | Minichromosome maintenance deficient (*S. cerevisiae*) 7 |
| 12 | 1859 | 307843 | ESTs |
| 13 | 809 | 810899 | CDC28 protein kinase 1 |
| 14 | 556 | 212198 | Tumor protein p53-binding protein, 2 |
| 15 | 3009 | 366647 | Butyrate response factor 1 (EGF-response factor 1) |
| 16 | 1443 | 566887 | Chromobox homolog 3 (drosophila HP1 gamma) |
| 17 | 1446 | 81331 | Fatty acid binding protein 5 (psoriasis-associated) |
| 18 | 1068 | 840702 | SELENOPHOSPHATE SYNTHETASE; Human selenium donor protein |
| 19 | 2893 | 32790 | MutS (*E. coli*) homolog 2 (colon cancer, nonpolyposis type 1) |
| 20 | 2699 | 44180 | Alpha-2-macroglobulin |

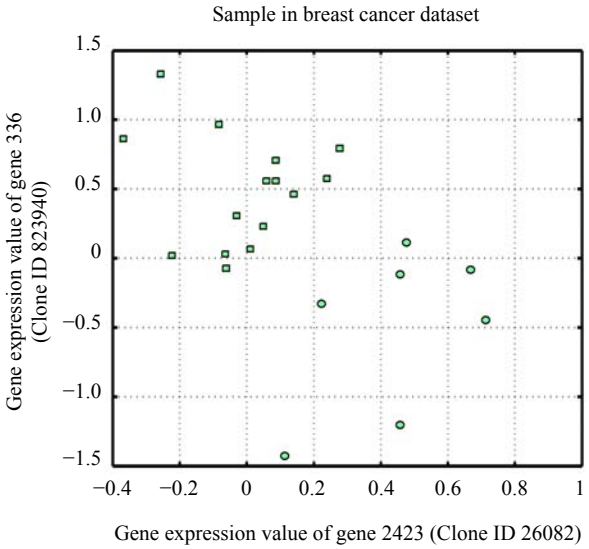


**Fig.3 Description of samples using the top 2 genes**

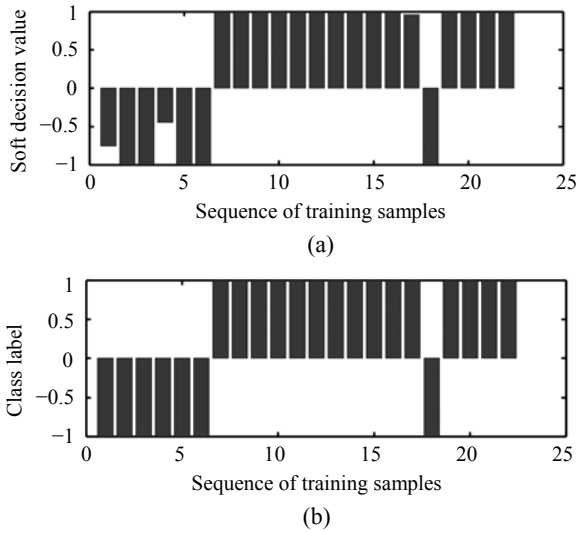


**Fig.4 Decision results of samples in breast cancer dataset (a) Soft decision value; (b) Class label**

cancer. Following the experimental setup in (Guyon *et al*., 2002), the data was split into a training set consisting of 38 samples of which 27 were ALL and 11 were AML, and a test set of 34 samples, i.e., 20 ALL and 14 AML. The data were preprocessed as recommended in (Dudoit *et al*., 2002): gene values were truncated from below at 100 and from above at 16000; genes maximum to minimum ratio of less than 5 or difference between the maximum and the minimum was less than 500 were excluded. Finally base-10 logarithm was applied to the 3571 remaining genes. Guyon *et al*.(2002) reported their data analysis revealed significant differences between the distribution of samples in the training set and in the test set, there are some important genes which are just as important to both training dataset and testing dataset.

**Table 2 Performance comparison of two gene ranking methods using linear SVM classifier on breast cancer dataset (22 samples)**

| Number of the top genes used | Linear SVM RFE with $C$=100 | | | | Gaussian kernel SVM RFE with model selected by GA | | | |
|---|---|---|---|---|---|---|---|---|
| | $V_{suc}$ | $V_{acc}$ | $V_{ext}$ | $V_{med}$ | $V_{suc}$ | $V_{acc}$ | $V_{ext}$ | $V_{med}$ |
| 32 | 1.00 | 1.00 | 0.52 | 0.85 | 1.00 | 1.00 | 0.20 | 0.84 |
| 16 | 0.91 | 0.73 | −0.20 | 0.61 | 1.00 | 1.00 | 0.36 | 0.81 |
| 8 | 0.55 | 0.00 | −1.00 | 0.10 | 1.00 | 1.00 | 0.50 | 0.92 |
| 4 | 0.36 | 0.00 | −1.00 | −0.44 | 1.00 | 1.00 | 0.59 | 0.95 |
| 2 | 0.50 | 0.00 | −1.00 | −0.45 | 0.95 | 0.91 | 0.31 | 0.91 |
| 1 | 0.68 | 0.00 | −1.00 | −0.42 | 0.64 | 0.00 | −0.35 | 0.41 |

**Table 3 Performance comparison of two gene ranking methods using Gaussian kernel SVM classifier on breast cancer dataset (22 samples)**

| Number of the top genes used | Linear SVM RFE with $C$=100 | | | | Gaussian kernel SVM RFE with model selected by GA | | | |
|---|---|---|---|---|---|---|---|---|
| | $V_{suc}$ | $V_{acc}$ | $V_{ext}$ | $V_{med}$ | $V_{suc}$ | $V_{acc}$ | $V_{ext}$ | $V_{med}$ |
| 32 | 1.00 | 1.00 | 0.58 | 0.88 | 1.00 | 1 | 0.48 | 0.88 |
| 16 | 0.95 | 0.95 | 0.26 | 0.85 | 1.00 | 1 | 0.60 | 0.90 |
| 8 | 0.95 | 0.95 | 0.29 | 0.81 | 1.00 | 1 | 0.29 | 0.84 |
| 4 | 0.77 | 0.00 | −1.00 | 0.42 | 1.00 | 1 | 0.57 | 0.88 |
| 2 | 0.82 | 0.00 | −1.00 | 0.55 | 1.00 | 1 | 0.51 | 0.86 |
| 1 | 0.82 | 0.00 | −1.00 | 0.47 | 0.91 | 0 | −0.17 | 0.70 |

**Table 4 Performance comparison of two gene ranking methods using linear SVM classifier on AML/ALL training dataset (38 samples)**

| Number of the top genes used | Linear SVM RFE with $C$=100 | | | | Gaussian kernel SVM RFE with model selected by GA | | | |
|---|---|---|---|---|---|---|---|---|
| | $V_{suc}$ | $V_{acc}$ | $V_{ext}$ | $V_{med}$ | $V_{suc}$ | $V_{acc}$ | $V_{ext}$ | $V_{med}$ |
| 32 | 1.00 | 1.00 | 0.52 | 0.85 | 1.00 | 1 | 0.62 | 0.94 |
| 16 | 0.91 | 0.73 | −0.20 | 0.61 | 1.00 | 1 | 0.65 | 0.95 |
| 8 | 0.55 | 0.00 | −1.00 | 0.10 | 1.00 | 1 | 0.67 | 0.95 |
| 4 | 0.36 | 0.00 | −1.00 | −0.44 | 1.00 | 1 | 0.90 | 0.99 |
| 2 | 0.50 | 0.00 | −1.00 | −0.45 | 1.00 | 1 | 0.75 | 0.98 |
| 1 | 0.68 | 0.00 | −1.00 | −0.42 | 0.71 | 0 | −1.00 | −0.38 |

**Table 5 Performance comparison of two gene ranking methods using Gaussian kernel SVM classifier on AML/ALL training dataset (38 samples)**

| Number of the top genes used | Linear SVM RFE with $C$=100 | | | | Gaussian kernel SVM RFE with model selected by GA | | | |
|---|---|---|---|---|---|---|---|---|
| | $V_{suc}$ | $V_{acc}$ | $V_{ext}$ | $V_{med}$ | $V_{suc}$ | $V_{acc}$ | $V_{ext}$ | $V_{med}$ |
| 32 | 1.00 | 1.00 | 0.46 | 0.92 | 1.00 | 1.00 | 0.70 | 0.92 |
| 16 | 0.97 | 0.97 | 0.20 | 0.91 | 1.00 | 1.00 | 0.70 | 0.94 |
| 8 | 0.97 | 0.95 | −0.01 | 0.88 | 1.00 | 1.00 | 0.58 | 0.94 |
| 4 | 0.89 | 0.00 | −0.60 | 0.80 | 1.00 | 1.00 | 0.34 | 0.92 |
| 2 | 0.95 | 0.00 | −0.55 | 0.85 | 1.00 | 1.00 | 0.08 | 0.90 |
| 1 | 0.82 | 0.00 | −1.00 | 0.58 | 0.97 | 0.97 | 0.10 | 0.91 |

**Table 6 Performance comparison of two gene ranking methods using linear SVM classifier on AML/ALL test dataset (34 samples)**

| Number of the top genes used | Linear SVM RFE with $C$=100 | | | | Gaussian kernel SVM RFE with model selected by GA | | | |
|---|---|---|---|---|---|---|---|---|
| | $V_{suc}$ | $V_{acc}$ | $V_{ext}$ | $V_{med}$ | $V_{suc}$ | $V_{acc}$ | $V_{ext}$ | $V_{med}$ |
| 32 | 0.94 | 0.91 | −0.11 | 0.85 | 0.97 | 0.91 | −0.12 | 0.90 |
| 16 | 0.82 | 0.00 | −1.00 | 0.61 | 0.94 | 0.88 | −0.18 | 0.82 |
| 8 | 0.74 | 0.00 | −1.00 | 0.41 | 0.97 | 0.94 | −0.08 | 0.86 |
| 4 | 0.71 | 0.00 | −1.00 | 0.20 | 0.97 | 0.94 | −0.13 | 0.93 |
| 2 | 0.59 | 0.00 | −1.00 | −0.29 | 0.94 | 0.94 | −0.48 | 0.91 |
| 1 | 0.59 | 0.00 | −1.00 | −0.39 | 0.59 | 0.00 | −1.00 | −0.39 |

**Table 7 Performance comparison of two gene ranking methods using Gaussian kernel SVM classifier on AML/ALL test dataset (34 samples)**

| Number of the top genes used | Linear SVM RFE with $C$=100 | | | | Gaussian kernel SVM RFE with model selected by GA | | | |
|---|---|---|---|---|---|---|---|---|
| | $V_{suc}$ | $V_{acc}$ | $V_{ext}$ | $V_{med}$ | $V_{suc}$ | $V_{acc}$ | $V_{ext}$ | $V_{med}$ |
| 32 | 0.94 | 0.91 | −0.11 | 0.86 | 0.97 | 0.91 | −0.14 | 0.89 |
| 16 | 0.97 | 0.79 | −0.17 | 0.87 | 0.94 | 0.91 | −0.16 | 0.82 |
| 8 | 0.97 | 0.82 | −0.42 | 0.87 | 0.97 | 0.94 | −0.07 | 0.07 |
| 4 | 0.97 | 0.00 | −0.33 | 0.90 | 0.97 | 0.94 | −0.08 | 0.94 |
| 2 | 0.79 | 0.00 | −0.63 | 0.62 | 0.94 | 0.94 | −0.42 | 0.92 |
| 1 | 0.76 | 0.00 | −0.87 | 0.42 | 0.91 | 0.00 | −1.00 | 0.83 |

If these genes are selected in the training dataset, a low error rate should be achieved on the testing dataset with these genes.

Two hundred important genes were pre-selected by $F$-test. One-hundred and forty-four populations were initialized, in which $C$ is 0.001 to 1000. For the parameters used in GA, $\ell$ was set at 3, $\eta$ at 0, $q$ at 0.05, $k_1$, $k_2$ at 0.8, and $k_3$, $k_4$ at 0.2. Similar to the treatment of breast cancer dataset, if the fitness value of the best population does not increase in four consecutive generations or the maximal time of 20 generations is reached, GA will be terminated. The optimized process is shown in Fig.5 showing that the GA converged after nine iterations, which the average fitness value is close to the best fitness value and that penalty parameter $C$ fluctuates in a big region between 100 and 400 in the first several generations, but in the final five generations, $C$ changes from 170 to 320. Note that the kernel width does not change much, and that the best fitness value is achieved with the smallest kernel width.

The top 20 genes are selected using the best model generated are listed in Table 8. The top gene, 4847 (Zyxin) is also considered the most important gene in (Guyon *et al*., 2002). And the gene 2354 (CCND3 CyclinD3), 6855 (TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)), 1834 (CD33 CD33 antigen (differentiation antigen)) and 5072 (Cytokeratin 17) are listed as important genes in (Lee *et al*., 2003). Samples in the training dataset and testing dataset are described in Fig.6 using the top 3 genes. No leave-one-out errors were found while only 1 test error was found in the testing dataset when Gaussian kernel SVM was used as classifier. The decision results of these two tests are illustrated in Fig.7. More detailed results are shown in Tables 4~7. In Table 4, linear SVM shows no leave-one-out errors occurred when the top 2 genes, top 4 genes, top 8 genes, top 16 genes and top 32 genes were used respectively; and in Table 6, linear SVM shows 1 test error occurred when the top 4 genes and top 8 genes were used respectively. In Table 5, Gaussian kernel SVM shows no leave-one-out errors occurred when the top 2 genes, top 4 genes, top 8 genes, top 16 genes and top 32 genes were used, and in Table 7, Gaussian kernel SVM shows that 1 test error occurred when the top 4 genes and top 8 genes were used.

Note that 1 leave-one-out error and 3 test errors

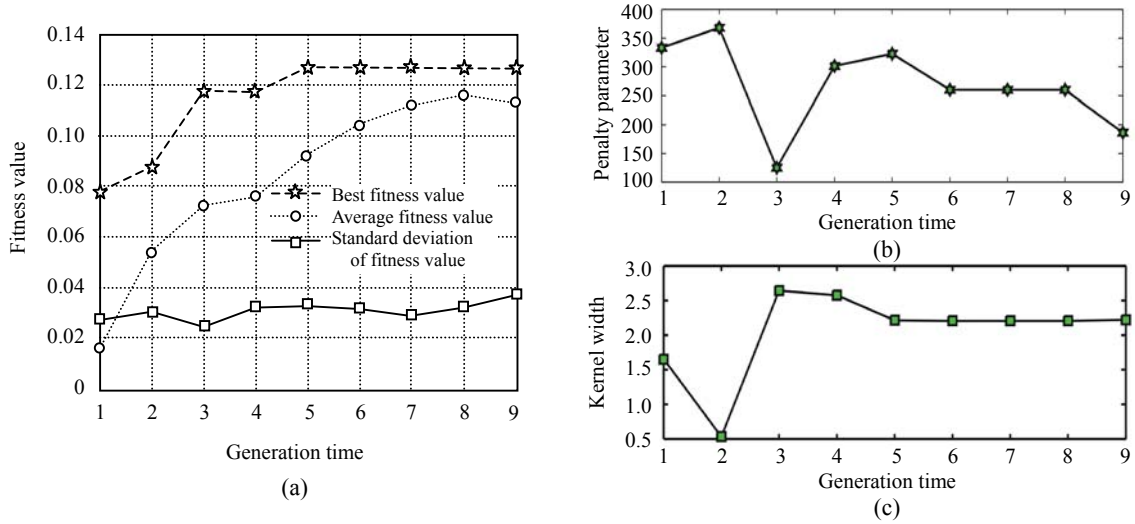


**Fig.5 Model optimization process for gene selection in AML/ALL dataset (a) Fitness value; (b) Penalty parameter; (c) Kernel width**

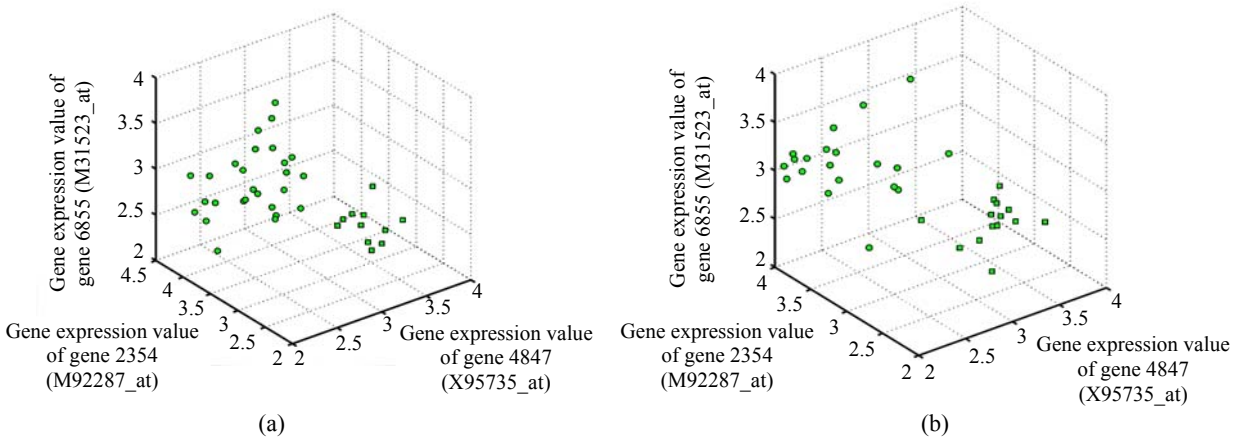


**Fig.6 Description of samples by the top 3 genes (a) Samples in training dataset by top 3 genes; (b) Samples in test dataset by top 3 genes**

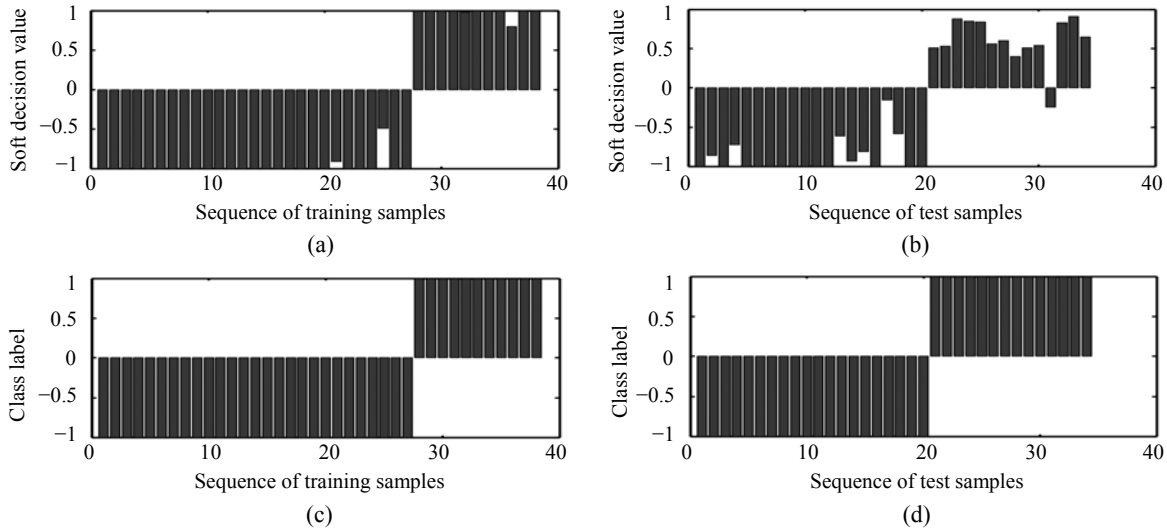


**Fig.7 Decision results of samples in AML/ALL dataset (a) Soft decision value of training samples and (b) test samples; (c) Class label of training samples and (d) test samples**

**Table 8 Top 20 genes ranked with optimized Gaussian kernel SVM RFE by GA on acute leukemia dataset**

| No. | Index No. | Accession number | Gene description |
|---|---|---|---|
| 1 | 4847 | X95735_at | Zyxin |
| 2 | 2354 | M92287_at | CCND3 Cyclin D3 |
| 3 | 6855 | M31523_at | TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47) |
| 4 | 5039 | Y12670_at | LEPR Leptin receptor |
| 5 | 2015 | M54995_at | PPBP Connective tissue activation peptide III |
| 6 | 1834 | M23197_at | CD33 CD33 antigen (differentiation antigen) |
| 7 | 1926 | M31166_at | "PTX3 Pentaxin-related gene, rapidly induced by IL-1 beta" |
| 8 | 5538 | D00097_s_at | "APCS Amyloid P component, serum" |
| 9 | 5358 | M85289_at | HSPG2 Heparan sulfate proteoglycan |
| 10 | 5072 | Z19574_rna1_at | Cytokeratin 17 |
| 11 | 5069 | Z18951_at | "CAV Caveolin, caveolae protein, 22000" |
| 12 | 3139 | U38864_at | Zinc-finger protein C2H2-150 mRNA |
| 13 | 1386 | L20591_at | ANX3 Annexin III (lipocortin III) |
| 14 | 1023 | HG544-HT544_at | Endothelial cell growth factor 1 |
| 15 | 387 | D38128_at | PTGIR Prostaglandin I2 (prostacyclin) receptor (IP) |
| 16 | 312 | D26308_at | NADPH-flavin reductase |
| 17 | 3877 | U85767_at | Myeloid progenitor inhibitory factor-1 MPIF-1 mRNA |
| 18 | 6128 | U43185_s_at | STAT5A Signal transducer and activator of transcription 5A |
| 19 | 3320 | U50136_rna1_at | Leukotriene C4 synthase (LTC4S) gene |
| 20 | 6218 | M27783_s_at | "ELA2 Elastatse 2, neutrophil" |

were found using top 4 genes ranked based on the whole dataset in (Guyon *et al*., 2002). In (Golub *et al*., 1999), 32 of 34 cases were correctly classified by using their top five genes.

**Discussion**

These algorithms were implemented with Matlab codes on an AMD 1800+ (1533 MHz) processor with 512 M memory (DDR 266 MHz). Using the RFE based on SVM with model selected by GA, the implementation yields a ranked list in about 5.5 h for the small round blue-cell tumors dataset (200 pre-selected genes and 22 samples) and 13.5 h for the acute leukemia dataset (200 pre-selected genes and 38 samples) in the experiments in this paper. Our experiments have revealed that $C$ has less effect on the ranked list, and that the fitness value depends largely on $\sigma^2$, which could be partly reflected in Fig.5. When $C$ is large enough (e.g. greater than 100), although it always fluctuates in a large area, the best fitness value nearly does not change. These phenomena occur frequently in our optimizing process. So, the selection of $\sigma^2$ is a key problem to gene ranking with Gaussian kernel SVM. Although the computational complexity of using Gaussian kernel SVM is so high, its performance is very satisfactory.

Tables 2~7 show that good classification performance will be achieved by the top 16 or top 32 genes selected by linear SVM RFE; and that using fewer genes (almost 4~8 genes) selected by our method, similar or better results will be achieved. When these two gene selection methods combined with a Gaussian kernel SVM classifier, the classifier performance is better than them combined with linear kernel SVM classifier.

## CONCLUSION

In this work, we studied the problem of gene selection in Gaussian kernel sphere with SVM. A machine learning method is proposed, which is RFE based on Gaussian kernel SVM with a model selected by GA. This method is a better alternative to the currently used common practice of selecting the apparently best parameters of Gaussian kernel SVM RFE. Based on certain fast implementation techniques, this method achieved satisfying results on the two hereditary breast cancer and acute leukemia datasets. The experimental results indicate that the

proposed methods perform well in selecting genes and achieve high classification accuracies with very few genes. Future work includes experimenting with application of this method to multi-classification problems and to other kernel methods. We envision that the non-linear classifiers are going to play an increasing important role in the analysis of cDNA microarray because of their superior performance in gene selection and cancer classification compared to existing methods.

## References


Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., *et al*., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**:503-511.

Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S., 2002. Choosing kernel parameters for support vector machines. *Machine Learning*, **46**:131-159.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines. Cambridge University Press, Cambridge.

Dudoit, S., Fridlyand, J., Speed, T.P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**:77-87.

Furlanello, C., Serafini, M., Merler, S., Jurman, G., 2003. An accelerated procedure for recursive feature ranking on microarray data. *Neural Networks*, **16**:641-648.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., *et al*., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**:531-537.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**:389-422.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Rafeld, M., *et al*., 2001. Gene expression profiles in hereditary breast cancer. *The New England Journal of Medicine*, **344**:539-548.

Houck, C., Joines, J., Kay, M., 1995. A Genetic Algorithm for Function Optimization: A Matlab Implementatio. NCSU-IE TR 95-09, North Carolina State University, USA.

Kim, S., Dougherty, E.R., Chen, Y., Sivakumar, K., Meltzer, P., Trent, J.M., Bittner, M., 2000. Multivariate measurement of gene expression relations. *Genomics*, **67**:201-209.

Kim, S., Dougherty, E.R., Barrea, J., Chen, Y., Bittner, M., Trent, J.M., 2002. Strong feature sets from small samples. *Journal of Computational Biology*, **9**:127-146.

Lee, K.E., Sha, N., Dougherty, E.R., Vannucci, M., Mallick, B.K., 2003. Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**:90-97.

Li, W., Yang, Y., 2002. How Many Genes are Needed for a Discriminant Microarray Data Analysis. *In*: Lin, S.M., Johnson, K.F. (Eds.), Methods of Microarray Data Analysis. Kluwer Academic, Boston, p.137-150.

Mao, Y., Zhou, X., Pi, D.Y., Wong, T.C., Sun, Y.X., 2004. Multi-class cancer classification by using fuzzy support vector machine and binary decision tree with gene selection. *Journal of Biomedicine and Biotechnology*, in Press.

Miettinen, K., Neittaanmaki, P., Makela, M.M., 1999. Evolutionary Algorithms in Engineering and Computer Science. Wiley, New York.

Shashua, A., Wolf, L., 2004. Kernel Feature Selection with Side Data using a Spectral Approach. Computer Vision−ECCV 2004: 8th European Conference on Computer Vision. Prague, Czech Republic, p.39-53.

Srinivas, M., Patnaik, L.M., 1994. Adaptive probabilities of crossover and mutation in genetic algorithm. *IEEE Trans. Syst. Man, Cybem.*, **24**(4):656-667.

Tabus, I., Astola, J., 2001. On the use of MDL principle in gene expression prediction. *J. Appl. Signal Process*, **4**:297-303.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory, 2nd Ed., Springer, New York.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., Vapnik, V., 2001. *In*: Leen, T.K., Dietterich, T.G., Tresp, V. (Eds.), Advances in Neural Information Processing Systems 13. MIT Press, Cambridge, MA, p.668-674.

Zhang, X., Wong, W., 2001. Recursive Sample Classification and Gene Selection Based on SVM: Method and Software Description. Technical Report, Department of Biostatistics, Harvard School of Public Health, USA.

Zhou, X., Wang, X., Dougherty, E.R., 2003a. Construction of genomic networks using mutual-information clustering and reversible-jump Markov-Chain-Monte-Carlo predictor design. *Signal Process*, **83**:745-761.

Zhou, X., Wang, X., Dougherty, E.R., 2003b. Binarization of microarray data based on a mixture model. *Molecular Cancer Therapeutics*, **2**:679-684.

Zhou, X., Wang, X., Dougherty, E.R., 2003c. Missing value estimation based on linear and nonlinear regression with Bayesian gene selection. *Bioinformatics*, **19**:2302-2307.

Zhou, X., Wang, X., Dougherty, E.R., 2004a. A Bayesian approach to nonlinear probit gene selection and classification. *Journal of Franklin Institute, Special Issue on Genomics, Signal Processing and Statistics*, **341**:137-156.

Zhou, X., Wang, X., Dougherty, E.R., 2004b. Nonlinear-probit gene classification using mutual-information and wavelet-based feature selection. *Biological Systems*, in Press.

Zhou, X., Wang, X., Dougherty, E.R., 2005. Gene selection using logistic regressions based on AIC, BIC and MDL criteria. *Journal of New Mathematics and Natural Computation*, **1**(1):129-145.