Selection for Feature Gene Subset in Microarray Expression Profiles Based on a Hybrid Algorithm Using SVM and GA

Wei Xiong, Chen Zhang, Chunguang Zhou, and Yanchun Liang

College of Computer Science and Technology, Jilin University, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Changchun 130012, China
ycliang@jlu.edu.cn

Abstract. It is an important subject to find feature genes from microarray expression profiles in the study of microarray technology. In this paper, a hybrid algorithm using SVM and GA is proposed. We first find a feature gene subset and filter most genes which are unrelated with diseases according to certain significant level, gene importance and classification efficiency by Least Square Support Vector Machine. Then we apply an improved genetic algorithm to carry out feature selection, in which the information entropy is used as a fitness function. At last, we apply the proposed feature selection algorithm to the two expression data sets of microarray, evaluate the feature gene subsets that are obtained in different conditions. Simulated results show that both good classification efficiency and the important genes which are related with diseases could be obtained by using the hybrid algorithm.

Keywords: feature selection, Support Vector Machine, genetic algorithm.

1 Introduction

The technology of microarray is a new technology with the development of life science and information technology. And the microarray is the most widely used technology in the fields of bioinformatics. The advent of microarray makes it possible to perform gene diagnosis and gene treatment.

Due to its low cost, high flux and high sensitivity, microarray is one of the important technologies for the study of functional genome, which is obviously better than the previous research model of single gene. But several challenging research tasks are largely overlooked because of the lack of an efficient analysis method.

Aiming at the above-mentioned problems, a feature selection method based on a SVM and GA hybrid algorithm is proposed to find a feature gene subset in this paper. At first, a feature gene subset is obtained and most genes which are unrelated with diseases according to certain significant level, gene importance and classification efficiency are filtered by Least Square Support Vector Machine. On the basis of it, we apply an improved genetic algorithm to carry out feature selection according to their contribution to classifying. In the proposed method, the crossover and mutation

G. Min et al. (Eds.): ISPA 2006 Ws, LNCS 4331, pp. 637 – 647, 2006.

operators in the genetic algorithm are improved such that the feature gene number of the subset could be controlled during the process of genetic operation, and the information entropy is used as separate criterion, and then the selected feature subset is evaluated by Support Vector Machine and the method of leave-one-out. We apply the proposed feature selection algorithm to the two expression data sets of microarray, evaluate the feature gene subsets that are obtained in different conditions. Comparisons of the simulated experimental results using the hybrid algorithm with those using other algorithms show the effectiveness of the proposed algorithm.

2 Model and Steps of the Hybrid Algorithm

The process of obtaining feature gene subset is made up of five steps:

(1) Perform preprocessing for the data of microarray. Use the Standard Deviation to do the data filtering and choose the genes with higher Standard Deviation and arrange the genes in order.

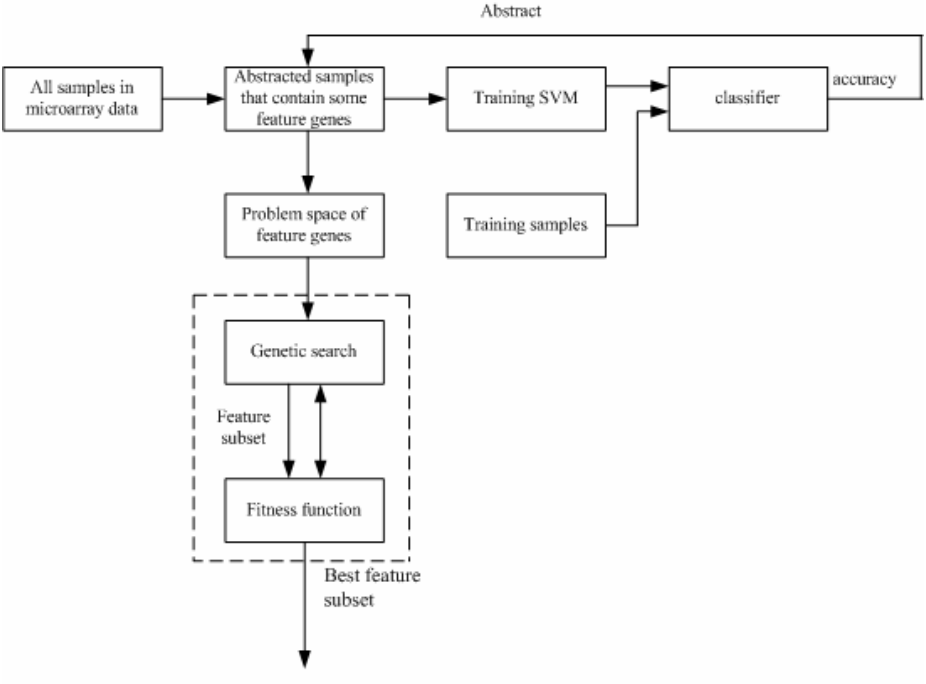

Fig. 1. Model of hybrid algorithm

(2) Use the Least Square Support Vector Machine (LSSVM) to construct a classifier. Firstly, select some samples to contain the training set and testing set, and then make selection to build up the certain quantity feature gene subset by the classified accuracy. This selecting and choosing process changes the gene number and compare the accuracy continuously.

(3) Compose the initial population with some individuals which have the feature genes selected by the LSSVM. And then make use of the improved genetic algorithm to perform genetic search. One individual is composed of some feature genes and a individual is a feature gene subset. Classify the initial training set with k-means clustering program, and then Classify the processed training set which has been taken part of genes that are not the feature genes out of the initial training set. The information entropy is used as the fitness function.
(4) Determine the individual whose fitness value is the largest one in the last generation. It is the best feature gene subset. The fitness value is the result computed by the information entropy.
(5) Evaluate the selected feature subset using SVM and the method of leave-one-out. Choose some of the samples to train and the others to test. Evaluate the result by the classified accuracy first and then observe the description and the function of the important feature gene selected. At last, compare the simulated experimental results using the hybrid algorithm with those using other algorithms, like Neural Network method.

The model of the feature gene subset selection is shown in Fig. 1.

3 Hybrid Algorithm of SVM and GA

3.1 Least Square Support Vector Machine

Support vector machine (SVM) was proposed by Vapnik and his research team based on statistical learning theory. The aim of SVM model is to construct the decision function takes the form [1, 2]:

$$f(x, w) = w^T \varphi(x) + b \tag{1}$$

where the nonlinear mapping $\varphi(\cdot)$ maps the input data into a higher dimensional feature space. In the LSSVM, the classification problem is formulated:

$$\min_{w,b,e} J(w,b,e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 \tag{2}$$

subject to the equality constraints

$$y_k [w^T \varphi(x_k) + b] = 1 - e_k, \quad k = 1,..., N \tag{3}$$

This corresponds to a form of ridge regression. The Lagrangian is given by

$$L(w,b,e;\alpha) = J(w,b,e) - \sum_{k=1}^{N} \alpha_k \{ y_k [w^T \varphi(x_k) + b] - 1 + e_k \} \tag{4}$$

with Lagrange multipliers α_k. The conditions for the optimality are

$$\begin{aligned}
&\frac{\partial L}{\partial W} = 0 \rightarrow w = \sum_{k=1}^{N} \alpha_k y_k \varphi(x_k) \\
&\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{k=1}^{N} \alpha_k y_k = 0 \\
&\frac{\partial L}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k, \quad k = 1,..., N \\
&\frac{\partial L}{\partial \alpha_k} = 0 \rightarrow y_k[w^T \varphi(x_k) + b] - 1 + e_k = 0, \quad k = 1,..., N
\end{aligned} \tag{5}$$

for $k=1,...,N$. After elimination of e_k and ω, the solution is given by the following set of linear equations

$$\begin{bmatrix} 0 & -Y^T \\ Y & ZZ^T + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ \vec{1} \end{bmatrix} \tag{6}$$

Where

$Z = [\varphi(x_1)^T y_1,..., \varphi(x_N)^T y_N]$, $y = [y_1,..., y_N]$, $\vec{1} = [1,...,1]$, $\alpha = [\alpha_1,..., \alpha_N]$

and Mercer condition

$$\Omega_{kl} = y_k y_l \varphi(x_k)^T \varphi(x_l) = y_k y_l \psi(x_k, x_l) \qquad k, l = 1,..., N \tag{7}$$

is applied. Hence, the classifier

$$f(x) = \text{sgn}\left(\sum_{k=1}^{N} \alpha_k^* y_k K(x \cdot x_k) + b^* \right) \tag{8}$$

is found by solving the linear set of Equations (6)-(7) instead of quadratic programming. Different kernel functions construct different SVM model. In this paper, we use three kernel functions to construct three different SVM model, Linear function, Polynomial function and Rbf function.

3.2 Improved Genetic Algorithm

Compose the initial population with some individuals which only have the genes selected by the LSSVM. The genetic operation is performed on these individuals. The operators in the improved genetic algorithm are fitness proportional selection operator, the improved crossover operator, and the improved mutation operator. The fitness function is based on the information entropy.

(1) Selection operator

The selection operation is based on the fitness, in which the fitness proportional selection operator is used.

(2) Improved crossover operator

It is demanded that the number of feature genes must keep fixed during the process of the genetic operation, so an improved crossover operator is developed. In the

crossover operation, make sure that the 0 positions and the 1 positions in one individual that changed by crossover operator are in the same quantity. For either of two individuals, cross over several 0 positions in one individual and 1 positions in the other individual when the corresponding positions in the two individuals are not the same. For example, in Fig. 2:

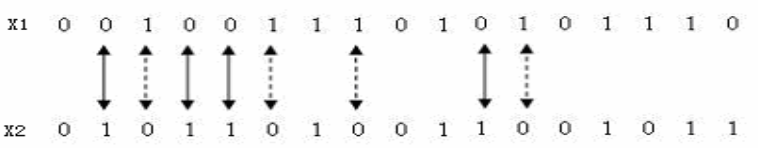

Fig. 2. The example of the improved crossover operator

(3) Improved mutation operator
For the same reason, an improved mutation operator is developed. When turning over a 1 position selected randomly in a individual, a 0 position is turned over.
(4) Fitness function
The construction of the fitness function is such an important part of genetic algorithm that it will affect the speed of convergence and the possibility of finding the best solution. In this paper, the fitness here corresponds to the separate criterion in feature extraction. The separate criterion based on the information entropy-gini impurity index is used in this paper [3].

Calculate the gini index of each category which is classified according to the current genetic individual. The fitness would be the reciprocal of the index sum. It can be written as：

$$f(G_i) = \frac{C}{\sum_{k=1}^{K} gini\ (\omega_k^i)} \tag{9}$$

Where C is an adjustment factor used to adjust the value of fitness, G_i represents the ith genetic individual after encoding, which is a feature gene subset. $\sum_{k=1}^{K} gini(\omega_k^i)$ is the sum of the index, in which：

$$gini\ (\omega_k^i) = \begin{cases} \frac{1}{2} H^2 \left[P(\omega_1 \mid \omega_k^i), P(\omega_2 \mid \omega_k^i), ..., P(\omega_n \mid \omega_k^i) \right] & n_k^i \neq 0 \\ 1 & n_k^i = 0 \end{cases} \tag{10}$$

When n_k^i, the number of samples which belong to category k according to the individual G_i, equals to 0, the diversity of categories is supposed to be the maximum, with the value 1, otherwise, it equals to the gini index, in which :

$$H(P(\omega_1 \mid x), P(\omega_2 \mid x),..., P(\omega_n \mid x)) = -\sum_{i=1}^{n} P(\omega_i \mid x) \log P(\omega_i \mid x) \quad (11)$$

square it under some condition, then we have

$$H^2(P(\omega_1 \mid x), P(\omega_2 \mid x),..., P(\omega_n \mid x)) = 2\left(1 - \sum_{i=1}^{n} P^2(\omega_i \mid x)\right) \quad (12)$$

where $P(\omega_i \mid x)$ represents the conditional probability of the sample belonging to the category ω_i under the condition *x*, it must satisfy that:

$$\sum_{j=1}^{n} P(\omega_j \mid x) = 1 \quad (13)$$

It represents the posterior probability that the samples which belong to the category *k* after reclassifying according to G_i belonged to the category *j* at the beginning. It satisfies apparently that:

$$\sum_{j=1}^{K} P(\omega_j \mid \omega_k^i) = 1 \quad (14)$$

So the information entropy-gini impurity index can be suitable for evaluating the classified result of feature gene subset.

The program is terminated if the difference of the mean fitness values of two generations is smaller than a threshold value. Furthermore, a maximum number of generations is set to make sure the algorithm be terminated actually. Using the improved genetic algorithm we get the best feature gene subset.

4 Experiment

4.1 Data Acquisition and Preprocessing

Microarray data consists of *p* genes and *n* DNA samples. The data can be described using a $p \times n$ matrix $X = [x_{ij}]$ where x_{ij} represents the expression data of the *i*th gene g_i on the *j*th sample X_j. A vector can be used to represent a sample:

$$X_j = \{x_{1j}, x_{2j},..., x_{pj}\}$$

Suppose that there are *p* genes in the microarray. The length of genetic string would be *p*, and a *p*-featured se is denoted by a binary vector *G*:

$$G = \{g_1, g_2,..., g_p\}$$

Every position in the string represents whether the relative gene would be involved in the subset. For example, a five-featured set could be $G = \{g_1, g_2, g_3, g_4, g_5\}$, a string {1 1 0 1 0} means that the subset is $G = \{g_1, g_2, g_4\}$.

Microarray expression profile has a lot of fuzzy data in it, which are no use for the classification. In order to obtain the right data for the next input used by SVM, microarray data need to be processed first. In the paper, a Standard Deviation process method is proposed to filter the expression data. The Standard Deviation is shown below:

$$S = \sqrt{\frac{1}{n}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2\right]} \tag{15}$$

4.2 Results of Experiments and Analysis

In order to evaluate the proposed hybrid algorithm, we perform experiments on two data sets, which are published in website. One is Colon data set and another is leukemia data set.

(1) Colon data set

The colon set is from Affymetrix Company, which is an array with 62 tissue samples of 65000 oligonucleotide genes (40 tumors and 22 normal tissues). In the experiment, we use the data with 2000 human genes picked by Alon et al, the expression data come from the website [4]: http://microarray.princeton.edu/oncology/affydata/.

For obtaining the right data for the next input, the microarray data need to be preprocessed. The results are shown in Fig. 3:

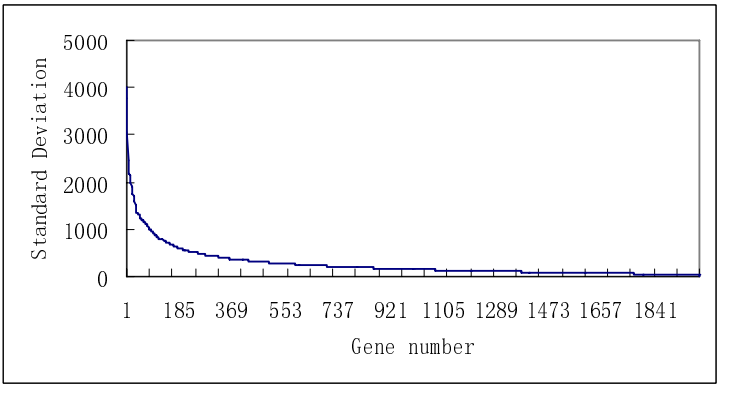

Fig. 3. Standard Deviation of 2000 genes in colon data set

Colon dataset contains two classes and it has 62 samples, among which 30 samples are for training and the remaining 32 samples are for testing. We use SVM with three kernel functions to evaluate the selected feature gene subset. The results are shown in Table 1.

The best result obtained using the hybrid algorithm is 20 genes. When the gene number is 20, the generation number is 800. The corresponding parameters in SVM need to be adjusted with gene subsets in different size. The result in this paper is also compared with the results obtained using other existing methods in references [5, 6]. The best accuracy obtained using a decision forest method combined with permutation method is 0.8767 when gene number is 39, and the best accuracy obtained using SVM method is 0.8390 when gene number is 20. The hybrid algorithm proposed with the best accuracy is 0.906 when the gene number is only 20.

Table 1. Evaluated results of subsets with different sizes

Gene Num	rbf	polynomial	linear
10	0.844	0.813	0.813
15	0.844	0.844	0.844
20	0.906	0.875	0.875
40	0.844	0.813	0.813
50	0.906	0.844	0.844

Because of the characteristics of gene selection, we not only need to get the good classification efficiency, but also obtain the important genes which are related with diseases. The results are shown in Table 2:

Table 2. Description of feature genes

Gene No.	**Sequence**	**Gene description**
U14973	Gene	Human ribosomal protein S29 mRNA, complete cds
T58861	3' UTR	60S RIBOSOMAL PROTEIN L30E (Kluyveromyces lactis)
Z22658	Gene	H.sapiens thrombin inhibitor mRNA
M26383	Gene	Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds
T74896	3' UTR	SERUM AMYLOID A PROTEIN PRECURSOR (HUMAN)
R01221	3' UTR	Human transcription factor TFIIA small subunit p12 mRNA, complete cds

Search in the bioinformatics database in networks, and compare with the results obtained using other experiments, it can be confirmed that these selected genes play an important role as a central hub for the gene network that maps to the underlying pathological complexity of colon cancer [3].

(2) Leukemia data set

Leukemia data set is from Golub's paper, which is composed with 72 tissue samples of 7129 leukemia genes (25 Acute Myeloid leukemia (AML) and 47 Acute lymphoblastic leukemia (ALL)). In the experiment, we use the mcroarray expression data from the website [7]: http://www.broad.mit.edu/cancer.

In order to obtain the right data for the next input using SVM, the microarray data need to be processed first, then we can select some feature genes from all of the 7129 genes. The Standard Deviation results are shown in Fig. 4.

The data set has 72 samples, among which 38 samples are for training and the remaining 34 samples are for testing. We also use the SVM with three kernel functions to evaluate the selected feature gene subset. The results are shown in Table 3.

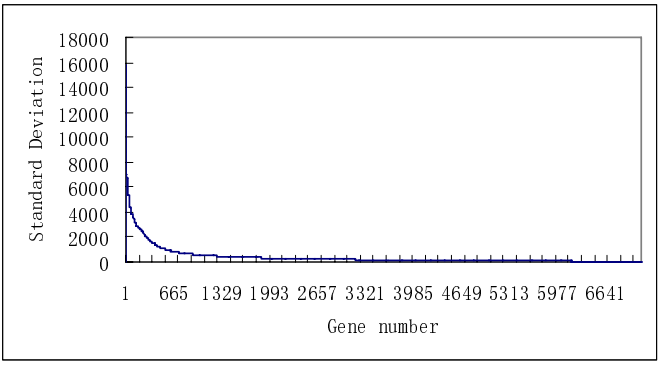

Fig. 4. Standard Deviation of 7129 genes in leukemia data set

Table 3. Evaluated results of subsets with different sizes

Gene num	rbf	polynomial	linear
10	0.941	0.941	0.941
15	0.971	0.971	0.971
20	0.971	0.941	0.941
40	0.971	0.971	0.971
50	0.971	0.971	0.971

The best result using the hybrid algorithm got is 15 genes. When the gene number is 15, the generation number is 1000, and the corresponding parameters in SVM need to be adjusted with gene subsets in different size.

The result in this paper is also compared with the results obtained using other methods in references [6-8]. The best accuracy obtained using a Neural Network method is 0.58, the best accuracy obtained using a SVM method is 0.971 when gene number is 20, and the best accuracy obtained by Golub et al is 0.853 when gene number is 50. In this paper, the proposed hybrid algorithm has the best accuracy of 0.971 when the gene number is only 15 compared with other methods.

Like the Colon dataset, we not only expect to get the good classification efficiency, but also obtain the important genes which are related with leukemia classification. The simulated experimental results are shown in Table 4.

Multiple lines of evidence from molecular biological studies imply that these genes are involved in leukemia development and progression, like M27891 [3]. Meanwhile, we also search in the bioinformatics database in networks, and compare the results with those obtained from other experiment, these feature genes concern the leukemia disease.

Table 4. Description of feature genes

Gene No.	Gene description
M27891	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
L25931	LBR Lamin B receptor
M38690	CD9 CD9 antigen
X95735	Zyxin
M31667	CYTOCHROME P450 IA2

We perform experiments on two data sets above, and the proposed hybrid algorithm shows the effectiveness compared with other methods.

5 Conclusions

Support Vector Machine has not only simple structure, but also better performances, especially better generalization ability. Meanwhile, GA has advantages in implicit parallelism, global optimum searching and simple operability. So the proposed hybrid algorithm based on SVM and GA is equal to select the feature genes, and numerical results show that good effectiveness of the proposed hybrid algorithm is obtained.

It can be seen from the two experiment results that the classified accuracy of selected feature gene subset is high, and the genes play an important role in the disease. And we know co-expressed genes may work in the same biological process. So hunting the important feature gene has made it possible and generated enormous interests to systematically derive biological pathways and networks. It is a challenging and meaningful task to find feature gene and build gene networks.

Acknowledgments. The authors are grateful to the support of the National Natural Science Foundation of China (60433020), the science-technology development project of Jilin Province of China (20050705-2), the doctoral funds of the National Education Ministry of China (20030183060), and “985” project of Jilin University.

References

1. Suykens JAK and Vandewalle J. Least Squares Support Vector Machines Classifiers. Neural Processing Letters. 9 (1999) 293-300.
2. Jiang JQ, Wu CG, and Liang YC. Multi-Category Classification by Least Squares Support Vector Regression. Lecture Notes in Computer Science. 3496 (2005) 863-868.
3. Li X, Rao SQ, Wang YD, et al. Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. Nucleic Acids Research. 9 (2004) 2685-2694.
4. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, and Levine AJ. Broad Patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA . 96 (1999) 6745-6750.
5. Lv SL, Wang QH, Li X, GU Z. Two feature gene recognition methods based on decision forest. China Journal of Bioinformatics. 3 (2004) 19-22.

6. Liu Q, Yang XT. Microarray Gene Expression Data Analysis Based on Support Vector Machine. Mini-Micro Systems. 3 (2005) 363-366.
7. Golub T R et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring .Science. 286 (1999) 531-537.
8. Toure A and Basu M. Application of neural network to gene expression data for cancer classification [C]. International Joint Conference on Neural Networks (IJCNN).1 (2001) 583-587.