Soft Comput (2008) 12:1039–1048
DOI 10.1007/s00500-007-0272-x

ORIGINAL PAPER

Gene selection using hybrid particle swarm optimization and genetic algorithm

Shutao Li · Xixian Wu · Mingkui Tan

Published online: 9 January 2008

Abstract Selecting high discriminative genes from gene expression data has become an important research. Not only can this improve the performance of cancer classification, but it can also cut down the cost of medical diagnoses when a large number of noisy, redundant genes are filtered. In this paper, a hybrid Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) method is used for gene selection, and Support Vector Machine (SVM) is adopted as the classifier. The proposed approach is tested on three benchmark gene expression datasets: Leukemia, Colon and breast cancer data. Experimental results show that the proposed method can reduce the dimensionality of the dataset, and confirm the most informative gene subset and improve classification accuracy.

Keywords Gene selection · Particle swarm optimization · Genetic algorithm · Support vector machine

1 Introduction

The rapidly developing DNA microarray technology can now measure large-scale gene expression data in a single experiment. However, gene expression data has characteristics of high-dimension, high-noise, and small-sample size. This gives rise to difficulties to a lot of classifiers. As the number of genes often exceeds tens of thousands, while the number of samples available is at most a few hundred, one of the main challenges in gene expression analysis is to determine genes which are relevant to a given cancer. But selecting high discriminative genes for microarray data remains a challenge (Tinker et al. 2006). Gene selection can improve the prediction accuracy of classifiers and save computational costs by using only discriminative genes with reduced dimensionality. More importantly, it enable the doctors to identify a small subset of biologically relevant genes with certain cancer as well as designing less expensive experiments by targeting only a small number of genes.

Several methods for informative gene selection have been proposed: TNoM (threshold number of misclassification) score (Ben-Dor et al. 2000), correlation metric (Golub et al. 1999; Furey et al. 2000), regression modeling approach (Thomas et al. 2001), mixture model approach (Pan 2002), and non-parametric tests (Troyanskaya et al. 2002; He 2004). However, these methods suffer from the deficiency that no correlation between the genes is considered in the selection procedure.

Support vector machines (SVM) (Vapnik 1995; Cristianini and Shawe-Taylor 1999) have demonstrated superior performance in classifying high dimension and sparse data. Several papers have reported good results on gene selection using SVM (Furey et al. 2000; Weston et al. 2000; Guyon et al. 2002; Zhang et al. 2005).

PSO and GA have been applied in feature selection. These are all evolutionary algorithm based on population (Shen et al. 2007; Li et al. 2001a,b; Ooi and Tan 2003; Peng et al. 2003). The PSO searches for the optimal solution by continually updating the particles' positions and velocities. GA finds the optimal solution by using chromosomes and GA operators including selection, crossover, and mutation. However, PSO and GA are easily trapped in local optimum in searching optimal solution. In order to overcome the local optimum problem, GA has been combined with PSO to find a better optimal solution (Shi et al. 2003).

In this paper, we present a gene selection method using combined PSO with GA using SVM. First, the Wilcoxon rank sum test (Deng et al. 2004) is used to preprocess the original gene expression data, and then the proposed hybrid PSO/GA

S. Li (✉) · X. Wu · M. Tan
College of Electrical and Information Engineering,
Hunan University, Changsha 410082, China
e-mail: shutao_li@yahoo.com.cn

is adopted to select the most important gene subsets using tenfold cross validation (CV) scheme. Finally, a classifier is trained based on the gene subset obtained from PSO/GA and used to predict the testing samples.

The organization of this paper is as follows. The proposed PSO/GA method is presented in detail in Sect. 2. In Sect. 3, experiments with three benchmark datasets are given. Conclusions of this paper are addressed in Sect. 4.

2 Method

2.1 Particle swarm optimization

Particle Swarm Optimization (PSO) originated from the simulation of social behavior of birds in a flock, which was developed by Kennedy and Eberhart (1995). In PSO, each particle flies in the search space with a velocity adjusted by its own flying memory and its companion's flying experience. All particles have fitness values which are decided by a fitness function. In this paper, the discrete binary version of PSO is used to select gene subset (Kennedy and Eberhart 1997). Its general steps are specified as below.

Firstly, discrete binary PSO uses fixed symbol serial in binary system as a particle in a swarm. One particle means one gene subset. The length of the particle is determined by the quantity of genes from preprocessing. Then all the particles search for the best solution in the solution space by making use of the best position of the particle. If it evolves into a certain generation, the procedure will be terminated and the generated particle (gene subset) is the best answer.

Each particle updates its own position and velocity according to formula (1) and (2) in every iteration.

$$v_{id}^{k+1} = \omega v_{id}^{k} + c_1\gamma_1\left(p_{id}^{k} - x_{id}^{k}\right) + c_2\gamma_2\left(p_{gd}^{k} - x_{id}^{k}\right) \quad (1)$$

$$x_{id}^{k+1} = \begin{cases} 1 & \mathrm{S}\left(v_{id}^{k+1}\right) > \mathrm{rand}(0,1) \\ 0 & \text{else} \end{cases} \quad (2)$$

where the $\mathrm{S}(v_{id})$ is the sigmoid function $\mathrm{S}(v_{id}) = 1/(1 + \exp(-v_{id}))$, $i = 1, 2, 3 \ldots m$, m is the number of particles in the swarm, v_{id}^{k} and x_{id}^{k} stand for the velocity and position of the ith particle of the kth iteration, respectively. p_{id}^{k} denotes the previously best position of particle i, p_{gd}^{k} denotes the global best position of the swarm. ω is the inertia weight, c_1 and c_2 are acceleration constants (the general value of c_1 and c_2 are in the interval [0 2]), γ_1 and γ_2 are random numbers in the range [0 1].

2.2 Genetic algorithm

Genetic algorithm (GA) is an adaptive optimization search algorithm simulating the evolutionary ideas of natural selection (Goldberg 1989). Its general steps can be specified as below. Like the PSO, the GA first randomly generates initial population, and then all the individual chromosomes are evaluated by a fitness function. The individuals search for the best solutions using the GA operators: selection, crossover, and mutation. The selection operator chooses the chromosomes with high adapting value from the current population. The crossover operator is used to combine two chromosomes to produce two new chromosomes called offspring. Mutation operator is to alter one or more gene values in a chromosome from its initial state. The process is repeated until the best fitness is satisfied or the last generation is arrived. A fitness function is used to evaluate the quality of each chromosome in the evaluation step. In the chromosome design, the binary coding system is used to represent the chromosome. Each bit of the chromosome represents a gene mask. The bit with value '1' indicates the gene is selected, and '0' represents the gene is discarded. All the genes with value '1' are selected and combined as a candidate gene subset. In this paper, the fitness of each chromosome (gene subset) is assessed by the classification accuracy of SVM. The 10-CV classification accuracy with the gene subset on the training samples is adopted. The higher the 10-CV classification accuracy, the better the gene subset is. The gene subset with the highest 10-CV classification accuracy is considered the optimal gene subset.

2.3 Hybrid PSO/GA

The main idea of hybrid PSO/GA algorithm is to integrate the GA operators into the PSO algorithm. Figure 1 shows the flow chart of the hybrid PSO/GA and details are presented as below:

Step 1. Generate initial population. Randomly generate M^*N initial population with binary system. M is the number of particles in a swarm, and N stands for the length of an individual (particle).

Step 2. Compute fitness. All the individuals are evaluated by a fitness function.

Step 3. Perform PSO operators. Each individual updates its position and velocity according to Eqs. 1 and 2.

Step 4. Judge termination. If an updated individual with new fitness cannot satisfy termination condition, go to step 5, otherwise the process output the final solution.

Step 5. Perform GA process.

Step 6. Compute fitness. This step is the same as step 2.

Step 7. Judge termination. Once the termination condition is met, output the final solution, otherwise go to step 3. The maximum number of iterations is considered as the termination criterion.

2.4 Support vector machine

Support Vector Machine (SVM) is specifically designed for two-class problems (Cristianini and Shawe-Taylor 1999).

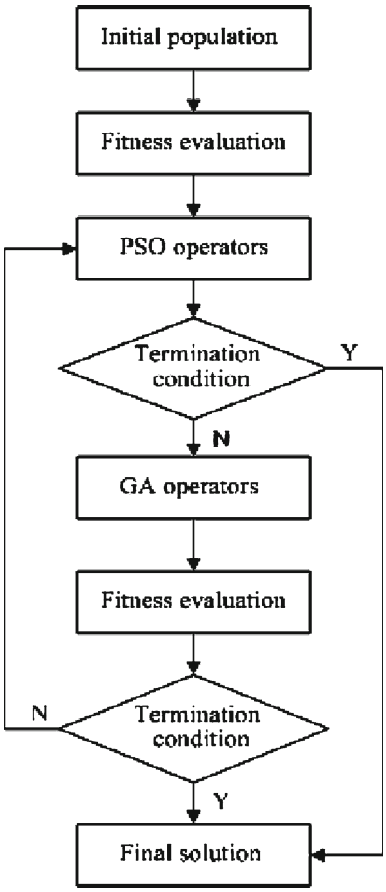

Fig. 1 Flow chart of hybrid PSO/GA

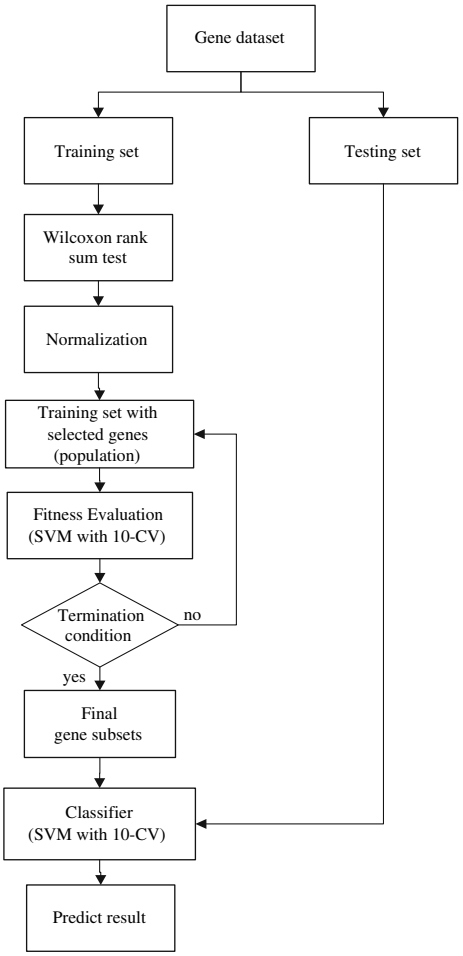

Fig. 2 Schematic illustration of the hybrid PSO/GA-based gene selection method

SVM can find a best hyperplane $(w^* x) + b = 0$ (w denotes normal vector of the plane and b is distance from plane to origin) between two classes of data. As for linearly separable case, after the classified plane divides the data into two classes, the margin between two classes data is $2/||w||$. The classifier is:

$$f(x) = \text{sgn}\left\{\sum_{i=1}^{n} a_i y_i (x_i \times x) + b\right\} \tag{3}$$

For the non-linear case, SVM will map the data in lower-dimensional space into a higher-dimensional space. The classifier is:

$$f(x) = \text{sgn}\left\{\sum_{i=1}^{n} a_i y_i K(x_i \times x) + b\right\} \tag{4}$$

where sgn{ } is the sign function, the a_i is Lagrange multiplier, x_i is a training sample, x is a sample to be classified, $K(x_i \times x)$ is the kernel function. Example kernel function includes Polynomial, Linear, and Radial basis function.

2.5 Gene selection

Figure 2 describes the basic procedure of a hybrid PSO/GA-based gene selection method. Details are described as follows:

Step 1: the gene expression data is preprocessed by the Wilcoxon rank sum test.

In this stage, all the training samples are firstly divided into training samples and testing samples using tenfold method. The tenfold procedure is: (1) the n samples are divided randomly into 10 subsets of (approximately) equal size; (2) 9 of the 10 subsets are used for gene selection and to train a classifier using the genes selected; (3) the remained subset is used to test the performance. This work should be done to each fold of the data. Secondly, the Wilcoxon rank sum test is performed on the training samples. In this process, a large number of redundant, noisy genes are filtered by Wilcoxon rank sum test. The statistics formula is as follow:

$$s(g) = \sum_{i \in N_0} \sum_{j \in N_1} I\left(\left(\mathbf{x}_j^{(g)} - \mathbf{x}_i^{(g)}\right) \leq 0\right) \tag{5}$$

Table 1 PSO/GA parameter settings

PSO/GA parameters	Leukemia	Colon	Breast cancer
Population	50	50	50
Individual length	40	40	40
Termination iterations	10	10	10
Inertia weight (ω)	0.9	0.9	0.9
Acceleration constants ($c_1 = c_2$)	2	2	2
Crossing rate	0.985	0.985	0.985
Mutation rate	0.05	0.05	0.05

Here I is the distinguishing function, if the logic expression $(\mathbf{x}_j^{(g)} - \mathbf{x}_i^{(g)}) \leqslant 0$ is true, I is 1, otherwise 0. $\mathbf{x}_i^{(g)}$ is the expression value of sample I in gene g, N_0 and N_1 stand for the number of samples belonging to the two classes, $s(g)$ represents the difference expression of one gene in the two classes. According to whether $s(g)$ reaches 0 or the maximum $N_0 \times N_1$, the corresponding gene is more important to classification. We use the expression below to evaluate importance of each gene.

$$q(g) = \max(s(g), N_0 \times N_1 - s(g)) \tag{6}$$

Each gene is evaluated and ranked according to equation (2), and the top N genes with the highest scores are selected as the new subset. In this paper, 40 top genes with the highest scores are selected as the crude gene subset. The corresponding genes in the testing samples are also selected at the same time.

Step 2: the final gene selection is performed using hybrid PSO/GA on the training samples.

Details of the hybrid PSO/GA have been described in Sect. 2.3. The individual (gene subset) is evaluated by a fitness function. The fitness function is the tenfold classification accuracy using SVM.

Step 3: the final most informative genes are selected, and then a classifier is trained using SVM based on the gene subset obtained from the proposed method. At last, the classifier is used to predict the testing samples.

3 Experiment results

3.1 Experimental setup

The proposed method is evaluated on three public data sets.

The leukemia dataset consists of 7129 genes and 72 samples from two different types of samples: acute lymphblastic leukemia (ALL) and acute myloid leukemia (AML). The training dataset contains 38 samples (27 ALL and 11 AML) while testing dataset consists of 34 samples (20 ALL and 14 AML) (Golub et al. 1999). The colon data set contains the expression of 2000 genes in 22 normal tissues and 40 colon tumor tissues (Alon et al. 1999). The breast cancer dataset consists of 7129 genes and 38 samples which contain 18 ER+ (estrogen receptor) samples and 20 ER-samples (West et al. 2001).

The hybrid PSO/GA parameters include PSO parameters and GA parameters. The PSO parameters are as follows: population (Swam), individual length (particle), inertia weight (ω), and acceleration constants ($c_1 = c_2$), the GA parameters contain Crossing rate and Mutation rate. In this paper, roulette wheel is used as the selection operator, double one-point crossover is adopted as the crossover operator, and the mutation operator is simple binary mutation. The Population is set to 50 on the three datasets, the crossing rate is set to 0.985 and the mutation rate is 0.05. The inertia weight and acceleration constants are constant by default. The final parameter settings are summarized in Table 1.

3.2 Experimental results

For the SVM, the Rbf kernel of is used, and the penalty factor C and Gamma are set at 2000, 0.0001, respectively.

The experiments are implemented using tenfold CV as discussed in Sect. 2.5. As the training set and testing set are changing under the tenfold CV strategy, the genes selected and the testing accuracy are different each time. Table 2 shows the testing accuracy and number of genes selected in 10 times on the three datasets. Here, Acc (%) is tenfold CV accuracy, Avg (N) is the average number of selected genes each time. From Table 2, the best classification accuracy on the leukemia data is 97.2% when an average of 18.7 genes are selected. The worst is 93.6% when 19.4 genes are selected. On the Colon data, the best testing accuracy is 91.9% with 18.0 genes and the worst is 87.1% with 15 genes. For the breast cancer data, the best classification result is 97.4% with 27.2 genes and the worst is 89.5% with 22.9 genes.

Considering the influence of the kernel types and parameters of SVM to the classification accuracy, Table 3 shows the results on the three datasets using different SVM kernel functions and parameters with optimal gene subset. The best ones are shown in bold. For the Leukemia data, the best classification result (97.2%) is obtained using Rbf kernel function

Table 2 Tenfold test accuracies with selected genes on three datasets

Running times	Data set					
	Leukemia		Colon		Breast cancer	
	Acc (%)	Avg (N)	Acc (%)	Avg (N)	Acc (%)	Avg (N)
1	97.2	22.5	90.3	16.3	97.4	30.2
2	95.8	22.0	91.9	18.0	94.7	27.4
3	93.6	19.7	87.1	15.0	92.1	27.5
4	94.4	19.6	88.7	16.0	92.1	28.1
5	97.2	18.7	88.7	17.8	89.5	22.9
6	94.4	21.6	88.7	16.1	92.1	27.8
7	93.6	19.4	87.1	18.2	92.1	26.2
8	94.4	23.0	90.3	16.7	94.7	27.1
9	94.4	21.7	87.1	18.5	92.1	25.2
10	95.8	19.2	87.1	15.7	97.4	27.2
Average	95.1	21.0	88.7	16.8	93.4	26.9

with $C = 3000$ and Gamma$= 0.0001$. For the Colon and Breast cancer data, the best results are also obtained with Rbf kernels. We conclude that the classification performance is more stable and effective using the Rbf kernel.

For further comparison, single PSO and single GA methods are also used for gene selection and compared to the proposed hybrid PSO/GA method. The corresponding parameter settings of the two algorithms are the same as the hybrid PSO/GA method. Table 4 shows the average classification accuracies (%) obtained by the three methods on 10 runs. The first number shows the average classification accuracy and the number in parenthesis is the average number of genes selected. From Table 4, we can see that the best classification results are 95.1% on Leukemia data using the hybrid PSO/GA methods, whereas, only 94.6% classification accuracy is obtained by single PSO or GA method. For the Colon data, an accuracy of 88.7% is obtained by the proposed method, which is also competitive to the results obtained by single PSO or GA method. For the breast cancer data, the best classification accuracy is 93.4% by hybrid PSO/GA, which is better than that of single PSO or GA. we can draw a conclusion that PSO algorithm with GA operators integrated in has better performance for gene selection compared to single PSO or GA.

In order to illustrate the good performance of the proposed gene selection method, Table 5 reports accuracy and number of genes selected by Naïve Bayes, C4.5, and SVM, and various gene selection algorithms, which all using tenfold CV evaluation (Ruiz et al. 2006). The first number stands for the classification accuracy while the number in parenthesis denotes the average number of genes selected. As shown in Table 5, for the Leukemia data, our proposed hybrid PSO/GA algorithm achieve the best accuracy using SVM and Naïve Bayes, which are 97.2% with an average of 18.7 and 18.9 genes, respectively. The following one is the Naïve Bayes classifier with FCBF algorithm which obtains a classification accuracy of 95.9% with an average of 45.8 genes. For the Colon data, the best classification accuracy is obtained by the proposed gene selection method with SVM classifier, which achieves 91.9% with an average of 18.0 genes. The combination of C4.5 classifier and FCBF method is the second highest one, which followed by our proposed method and BIRSw with Naïve Bayes. As shown in Table 5, for Leukemia data, the proposed gene selection method with NB and SVM achieves the best results. For Colon data, the proposed method with SVM is the best one.

3.3 Further analysis of the experimental results

We now do some further analysis using the selected genes by previous tenfold CV. Firstly, the appearances of the genes are counted. Figures 3, 4 and 5 show the frequencies of the selected gene appearances on the three datasets. Horizontal axis denotes the index of genes occurring in the processing, and vertical axis stands for the number of appearances of the corresponding genes.

Then, the genes are ranked based on the number of appearances on the three datasets. At last, the top K genes with the highest appearances are selected and tested using a SVM with tenfold CV on the samples. Figure 6 shows the tenfold classification accuracy with the selected top K genes on the three datasets. The horizontal axis denotes the top K genes selected. Here the upper bond of K is set at 30, and the vertical axis stands for the corresponding tenfold classification accuracy using SVM. In the experiment, the Rbf kernel is adopted, the penalty factor C is set to 2000 and Gamma is set to 0.001 on three datasets.

From Fig. 5, we can see that the tenfold accuracy of 97.7% is obtained with the top three genes on the Leukemia dataset. For the Colon dataset, the tenfold classification accuracy of

Table 3 The testing accuracies (%) with different SVM parameters

Data set	SVM kernel	SVM parameter (C)	Selected genes	Tenfold accuracy (%)
Leukemia	Linear	1000	23.5	95.8
		2000	24.5	94.4
		3000	23.8	95.8
	Polynomial (Degree = 2)	1000	20.5	94.4
		2000	25.9	93.1
		3000	22.7	94.4
	Rbf (Gamma = 0.001)	1000	23.8	93.1
		2000	23.0	95.8
		3000	19.8	97.2
	Rbf (Gamma = 0.0001)	1000	19.9	94.4
		2000	23.5	**97.2**
		3000	22.5	94.4
Colon	Linear	1000	15.9	74.2
		2000	15.6	74.2
		3000	16.7	79.3
	Polynomial (Degree = 2)	1000	16.6	82.3
		2000	17.2	83.9
		3000	16.6	75.8
	Rbf (Gamma = 0.001)	1000	11.8	85.5
		2000	19.1	85.5
		3000	16.2	82.3
	Rbf (Gamma = 0.0001)	1000	19.4	88.7
		2000	17.9	**90.3**
		3000	18.1	88.7
Breast cancer	Linear	1000	28.2	92.1
		2000	28.6	94.7
		3000	28.4	92.1
	Polynomial (Degree = 2)	1000	28.6	86.8
		2000	24.2	92.1
		3000	27.4	89.5
	Rbf (Gamma = 0.001)	1000	28.0	94.7
		2000	28.9	94.7
		3000	28.2	92.1
	Rbf (Gamma = 0.0001)	1000	25.1	92.1
		2000	27.4	94.7
		3000	27.1	**97.3**

Table 4 Testing accuracies (%) obtained by the three methods on three datasets

Method	Leukemia	Colon	Breast cancer
Single PSO	94.6(22.3)	87.1(19.8)	91.8(29.4)
Single GA	94.6(23.1)	87.1(17.5)	91.6(28.9)
Hybrid PSO/GA	95.1(21.0)	88.7(16.3)	93.4(26.9)

91.9% is obtained when the top nine genes are selected. For the breast cancer data, the tenfold classification accuracy of 100% is obtained with the top 11 genes.

Tables 6, 7 and 8 list the top 20 optimal genes selected by the proposed method and their corresponding indices in the three original datasets.

Figures 7, 8 and 9 show the expression values of the selected genes of the three datasets. The columns denote the genes, and the rows represent the corresponding expression levels. In Fig. 7 the left 25 columns are AML samples and the right 47 columns are ALL samples for Leukmekia data. For the Colon data, as shown in Fig. 8, the left 22 columns are normal tissue and the right 40 columns are tumor tissues. In Fig. 9, the left 18 columns are ER+ patients and the right

Table 5 Testing accuracies (%) obtained by the various methods as reported in the literatures

Classifier	Selection algorithm	Leukemia	Colon
NB	BIRSw (Ruiz et al. 2006)	93.4(2.5)	85.5(3.5)
	FCBF (Ruiz et al. 2006; Yu and Liu 2004)	95.9(45.8)	77.6(14.6)
	SFw (Yu and Liu 2004)	87.3(3.2)	84.1(5.9)
	CFSsf (Hall 2000)	91.4(40.3)	82.6(22.1)
	FOCUSsf (Almuallim and Dietterich 1994)	84.8(2.4)	77.1(4.6)
	Hybrid PSO/GA	**97.2(18.9)**	85.5(12.9)
C4.5	BIRSw (Ruiz et al. 2006)	88.6(1.2)	83.8 (2.9)
	FCBF (Ruiz et al. 2006; Yu and Liu 2004)	83.2(45.8)	88.3(14.6)
	SFw (Yu and Liu 2004)	87.3(1.6)	80.7(3.3)
	CFSsf (Hall 2000)	84.8(40.3)	86.9(22.1)
	FOCUSsf (Almuallim and Dietterich 1994)	88.9(2.4)	79.1(4.6)
	Hybrid PSO/GA	91.7(20.1)	83.9(15.1)
SVM	Hybrid PSO/GA	**97.2(18.7)**	**91.9(18.0)**

Table 6 The top 20 important genes selected from the Leukemia cancer dataset

Rank	Index	Gene description
1	804	Macmarcks
2	1685	Terminal transferase mRNA
3	2121	CTSD Cathepsin D (lysosomal aspartyl protease)
4	760	CYSTATIN A
5	1144	SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)
6	1630	Inducible protein mRNA
7	1745	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
8	1779	MPO Myeloperoxidase
9	1953	Fc-epsilon-receptor gamma-chain mRNA
10	2402	Azurocidin gene
11	4229	SPI1 Spleen focus forming virus (SFFV) proviral integration oncogene spi1
12	4366	ARHG Ras homolog gene family, member G (rho G)
13	4377	ME491 gene extracted from H.sapiens gene for Me491/CD63 antigen
14	6376	PFC Properdin P factor, complement
15	1829	PPGB Protective protein for beta-galactosidase (galactosialidosis)
16	1834	CD33 CD33 antigen (differentiation antigen)
17	1882	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
18	2288	DF D component of complement (adipsin)
19	4377	ME491 gene extracted from H.sapiens gene for Me491/CD63 antigen
20	6855	TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)

20 columns are ER- patients. From Figs. 7, 8 and 9, we can see that the selected genes can discriminate the three datasets into two classes.

4 Conclusions

In this paper a hybrid PSO/GA method is proposed for gene selection and tested on three public gene datasets. Firstly, all the training samples are divided into training samples and testing sample using tenfold method. Secondly, the Wilcoxon rank sum test is adopted to find a crude gene subset on the training samples. Thirdly, the proposed hybrid PSO/GA method is used to perform the final gene selection based on the crude gene subset. In the proposed method, the GA operators are integrated into the PSO algorithm to improve the gene selection performance. Then, a classifier is trained on the optimal gene subset and used to predict the testing samples.

Table 7 The top 20 important genes selected from the Colon dataset

Rank	Index	Gene description
1	625	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein
2	780	MACROPHAGE MIGRATION INHIBITORY FACTOR (HUMAN)
3	1843	GELSOLIN PRECURSOR, PLASMA (HUMAN)
4	377	H.sapiens mRNA for GCAP-II/uroguanylin precursor
5	964	NUCLEOSIDE DIPHOSPHATE KINASE A (HUMAN)
6	267	Human cysteine-rich protein (CRP) gene, exons 5 and 6
7	391	Human mRNA (KIAA0069) for ORF (novel protein), partial cds
8	493	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)
9	249	Human desmin gene, complete cds
10	739	TROPOMYOSIN ALPHA CHAIN, SMOOTH MUSCLE (HUMAN)
11	137	Human mRNA (KIAA0027) for ORF, partial cds
12	824	H.sapiens gene for chemokine HCC-1
13	897	COMPLEMENT FACTOR D PRECURSOR (Homo sapiens)
14	365	Human hmgI mRNA for high mobility group protein Y
15	1042	P03001 TRANSCRIPTION FACTOR IIIA
16	1582	H.sapiens mRNA for p cadherin
17	1771	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds
18	245	Human cysteine-rich protein (CRP) gene, exons 5 and 6
19	67	CYSTATIN C PRECURSOR (HUMAN)
20	14	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (HUMAN).

Table 8 The top 20 important genes selected from the breast cancer dataset

Rank	Index	Gene description
1	4445	Human breast cancer, estrogen regulated LIV-1 protein (LIV-1) mRNA, partial cds
2	5914	Human hGATA3 mRNA for trans-acting T-cell specific transcription factor
3	2730	Human steroid 5-alpha-reductase mRNA, complete cds
4	495	Human mRNA for KIAA0075 gene, partial cds
5	5782	H.sapiens pS2 protein gene
6	5639	Human mRNA for cytochrome c oxidase subunit VIc
7	5188	Human clone 23948 mRNA sequence
8	3272	Human glucokinase (GCK) mRNA, complete cds
9	1505	Human microtubule-associated protein tau mRNA, complete cds
10	5953	Human rearranged mRNA for glutamine synthase
11	5859	H.sapiens GATA-3 mRNA
12	4414	Human hepatocyte nuclear factor-3 alpha (HNF-3 alpha) mRNA, complete cds
13	1208	Nuclear Factor Nf-Il6
14	6419	H.sapiens LU gene for Lutheran blood group glycoprotein
15	6247	H.sapiens ARSE mRNA
16	5524	Human mRNA for estrogen receptor
17	5433	Homo sapiens Nedd-4-like ubiquitin-protein ligase WWP1 mRNA, partial cds. /gb=U96113 /ntype=RNA
18	4084	Human fructose-1,6-biphosphatase (FBP1) gene
19	3598	GCN5-like 1=GCN5 homolog/putative regulator of transcriptional activation {clone GCN5L1} [human, mRNA, 545 nt]
20	3406	Homo sapiens (pp21) mRNA, complete cds

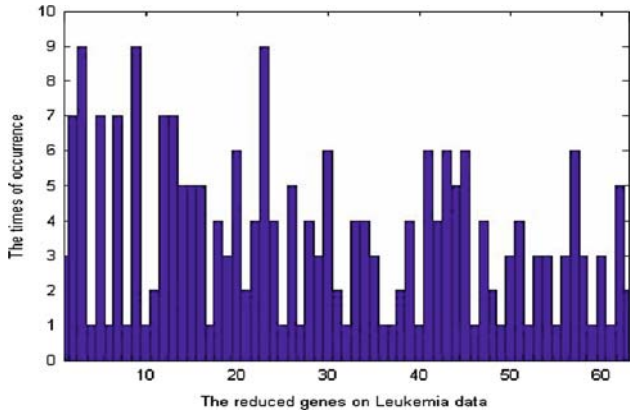

Fig. 3 The frequency of gene appearance on Leukemia data

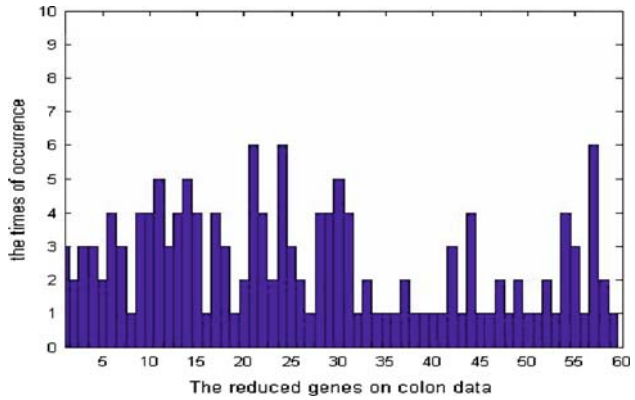

Fig. 4 The frequency of gene appearance on Colon data

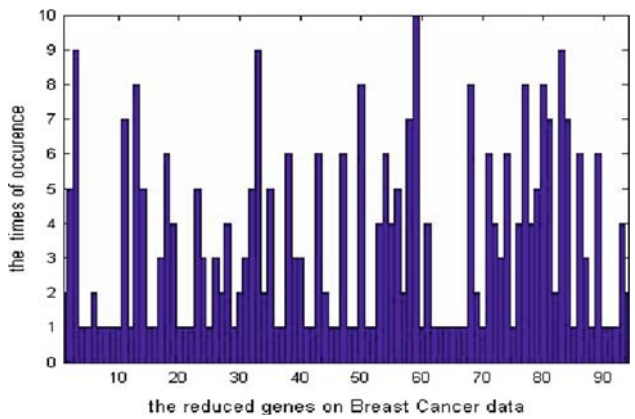

Fig. 5 The frequency of gene appearance on breast cancer data

Further analysis is given to confirm the genes selected. Experimental results with the Leukemia dataset and the Colon dataset suggest that the proposed strategy can reduce the dimensionality of the dataset, improve the classification accuracy and confirm the most informative gene subset for classification. In the near future, the optimization of parameters for SVM should be studied.

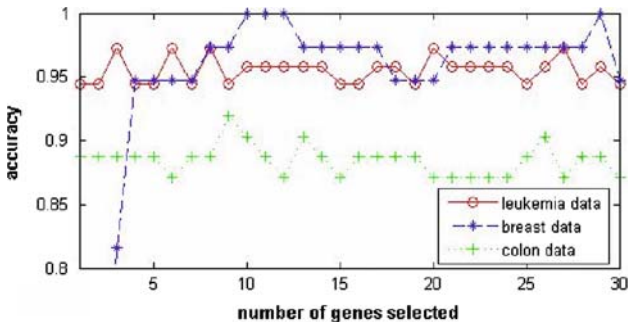

Fig. 6 The tenfold CV testing accuracy of top K genes selected on the three datasets

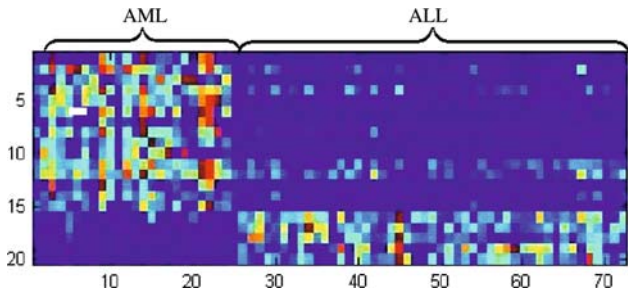

Fig. 7 The expression values of top 20 genes selected in two samples of Leukemia data

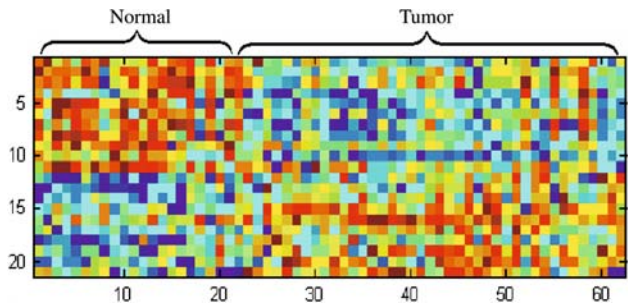

Fig. 8 The expression values of top 20 genes selected in two samples of Colon data

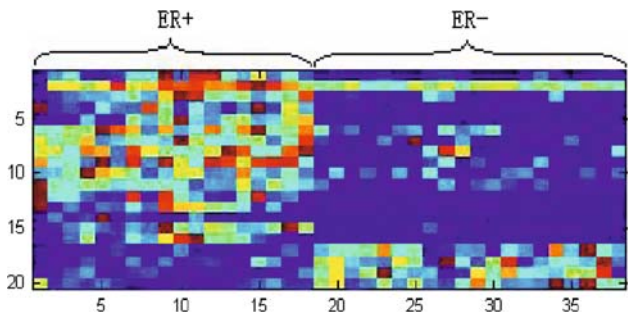

Fig. 9 The expression values of top 20 genes selected in two samples of breast cancer data

Acknowledgments The authors would like to thank the anonymous reviewers for their detailed review, valuable comments and constructive suggestions. This paper is supported by the Program for New Century Excellent Talents in University(NECT-2005), and the Excellent Youth Foundation of Hunan Province (06JJ1010).

References

Almuallim H, Dietterich T (1994) Learning boolean concepts in the presence of many irrelevant features. Artif Intell 69(1–2):279–305

Alon U, Barkai U, Notterman DA et al (1999) Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci 96:6745–6750

Ben-Dor A, Bruhn L, Friedman N et al (2000) Tissue classification with gene expression profiles. J Comput Biol 7:559–583

Cristianini N, Shawe-Taylor J (1999) An introduction to SVM. Cambridge University Press, Cambridge

Deng L, Pei J, Ma J et al (2004) A rank sum test method for informative gene discovery. In: Kim W, Kohavi R, Gehrke J et al (eds) Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 410–490

Furey TS, Cristianini N, Duffy N et al (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16:906–914

Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Addison-Wesley, New York

Golub T, Slonim D, Tamayo P et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 28:531–537

Guyon I, Weston J, Barnhill S et al (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422

Hall M (2000) Correlation-based feature selection for discrete and numeric class machine learning. In: 17th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA

He W (2004) A spline function approach for detecting differentially expressed genes in microarray data analysis. Bioinformatics 20:2954–2963

Kennedy J, Eberhart RC (1995) Particle swarm optimization. In Proceedings of the IEEE International Conference on Neural Networks, pp 1942–1948

Kennedy J, Eberhart RC (1997) A discrete binary version of the particle swarm algorithm. In Proceedings of the IEEE International Conference on Systems. Man, and Cybernetics, pp 4104–4109

Li L, Darden TA, Weingberg CR et al (2001a) Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. Comb Chem High Throughput Screen 4:727–739

Li L, Weinberg CR, Darden TA et al (2001b) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 17:1131–1142

Ooi CH, Tan P (2003) Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics 19:37–44

Pan W (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated Microarray experiments. Bioinformatics 18:546–554

Peng S, Xu Q, Ling XB et al (2003) Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. FEBS Lett 555:358–362

Ruiz R, Riquelme JC, Aguilar-Ruiz JS (2006) Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recognit 39(12):2383–2392

Shen Q, Shi WM, Kong W et al (2007) A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification. Talanta 71:1679–1683

Shi XH, Lu YH, Zhou CG et al (2003) Hybrid evolutionary algorithms based on pso and ga. In: Sarker R, Reynolds R, Abbass H et al (eds) Proceeding of IEEE Congress on Evolutionary computation, pp 2393–2399

Thomas JG, Olson JM, Tapscott SJ et al (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res 11:1227–1236

Tinker AV, Boussioutas A, Bowtell DDL (2006) The challenges of gene expression microarrays for the study of human cancer. Cancer Cell 9:333–339

Troyanskaya OG, Garber ME, Brown PO et al (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 18:1454–1461

Vapnik V (1995) The nature of statistical learning theory. Springer, New York

West M, Blanchette C, Dressman H et al (2001) Predicting the clinical status of human breast cancer using gene expression profiles. Proc Natl Acad Sci 98:11462–11467

Weston J, Mukherjee S, Chapelle O et al (2000) Feature selection for SVMs. Adv Neural Inf Process Syst 13:668–674

Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224

Zhang H, Ahn J, Lin X et al (2005) Gene selection using support vector machines with non-convex penalty. Bioinformatics 22:88–95