Available online at www.sciencedirect.com
ScienceDirect
BioSystems 90 (2007) 78–86
BioSystems
www.elsevier.com/locate/biosystems

Selecting a minimal number of relevant genes from microarray data to design accurate tissue classifiers

Hui-Ling Huang [a], Chong-Cheng Lee [b], Shinn-Ying Ho [c,d,*]

[a] *Department of Information Management, Jin-Wen Institute of Technology, Hsin-Tien 231, Taiwan*
[b] *Institute of Information Engineering and Computer Science, Feng Chia University, Taichun 407, Taiwan*
[c] *Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan*
[d] *Institute of Bioinformatics, National Chiao Tung University, Hsinchu 300, Taiwan*

Received 23 October 2005; received in revised form 28 April 2006; accepted 5 July 2006

Abstract

It is essential to select a minimal number of relevant genes from microarray data while maximizing classification accuracy for the development of inexpensive diagnostic tests. However, it is intractable to simultaneously optimize gene selection and classification accuracy that is a large parameter optimization problem. We propose an efficient evolutionary approach to gene selection from microarray data which can be combined with the optimal design of various multiclass classifiers. The proposed method (named GeneSelect) consists of three parts which are fully cooperated: an efficient encoding scheme of candidate solutions, a generalized fitness function, and an intelligent genetic algorithm (IGA). An existing hybrid approach based on genetic algorithm and maximum likelihood classification (GA/MLHD) is proposed to select a small number of relevant genes for accurate classification of samples. To evaluate the performance of GeneSelect, the gene selection is combined with the same maximum likelihood classification (named IGA/MLHD) for convenient comparisons. The performance of IGA/MLHD is applied to 11 cancer-related human gene expression datasets. The simulation results show that IGA/MLHD is superior to GA/MLHD in terms of the number of selected genes, classification accuracy, and robustness of selected genes and accuracy.

Keywords: Classification; Feature selection; Genetic algorithm; Maximum likelihood; Microarray

1. Introduction

Microarray is a useful technique for measuring expression data of thousands of genes simultaneously. The prediction of the diagnostic category of a tissue sample from its expression array phenotype using the microarray data from tissues in identified categories is known as classification. The samples are usually the experiments and the categories are the types of tissue samples. The number of genes is usually much greater than the number of tissue samples available, and only a small subset of the genes is relevant in distinguishing different classes. Therefore, one major challenge in designing the accurate classifiers using microarray data is to identify the optimal subset of relevant genes which is known as gene selection, corresponding to feature selection in the field of pattern classification.

Li et al. (2004) proposed a comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Most feature selection methods are univariate that each candidate relevant gene is considered individually. Recently, multivariate methods considering multiple genes simultaneously are rapidly gaining attention due

* Corresponding author. Tel.: +886 35131405; fax: +886 35729288.
E-mail address: syho@mail.nctu.edu.tw (S.-Y. Ho).

0303-2647/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2006.07.002

to the effectiveness of reducing number of relevant genes. Lee et al. (2003) used a hierarchical Bayesian model which used a Markov Chain Monte Carlo-based stochastic search algorithm to discover relevant genes. Yeung and Bumgarner (2003) proposed a multivariate gene selection method by eliminating highly corrected genes to reduce the number of relevant genes.

To advance the classification performance using a small number of features, it is better to take feature selection and classifier design into account simultaneously (Ho et al., 2002). Genetic algorithm (GA; Goldberg, 1989) is a randomized search and optimization technique that simulates the natural evolution by an iterative computation process. GA can consider multiple interacting attributes simultaneously rather than considering a single attribute at a time. Li et al. (2001) proposed a hybrid method of GA-based gene selection and k-nearest neighbor classifier to assess the importance of genes for classification. Ooi and Tan (2003) proposed an efficient hybrid approach based on GA and maximum likelihood classification (GA/MLHD).

In this paper, we propose an efficient evolutionary approach to gene selection from microarray data which can be combined with the optimal design of various multiclass classifiers, such as support vector machine, naive Bayes, k-nearest neighbor, and decision tree. The proposed method (named GeneSelect) consists of three parts which are fully cooperated: an efficient encoding scheme of candidate solutions, a generalized fitness function, and an intelligent genetic algorithm (IGA). GeneSelect aims to determine a minimal number of relevant genes and identify these genes, while maximizing classification accuracy simultaneously.

To evaluate the performance of GeneSelect, the gene selection is combined with the maximum likelihood classification for convenient comparisons (named IGA/MLHD). The performance of the proposed IGA/MLHD is evaluated using 11 cancer-related human gene expression datasets. The simulation results show that IGA/MLHD is superior to GA/MLHD in terms of the number of selected genes, classification accuracy, and robustness of selected genes and accuracy, especially for the datasets having numerous categories.

2. Methods

GeneSelect is a generalized gene selection method for microarray analysis as well as an efficient feature selection method in bioinformatics. The investigated gene selection problem is to identify a minimal number G from N genes of microarray data, while maximizing classification accuracy R. It is intractable to simultaneously optimize the two objectives (minimizing G and maximizing R) because of $G \ll N$. Essentially, the problem is a bi-objective binary combinatorial optimization problem with a search space $C(N, G)$ (Ho et al., 2004a). It is desirable to design a specific GA for efficiently searching the space $C(N, G)$ in a limited amount of computation time, considering the property $G \ll N$. The proposed GeneSelect consists of a novel solution encoding scheme, an associated fitness function, and IGA with an efficient crossover operation for maintaining feasibility of candidate solutions, described below.

2.1. Encoding of candidate solutions

In the GA of GeneSelect, a candidate solution used to identify the subset of selected genes for optimizing a fitness function (or conventional objective function) is encoded, named GA-chromosome. It is impractical to directly encode N decision variables into a GA-chromosome using an N-bit string because N is too large (generally, $N \in [2000, 20000]$). To cope with difficulties of this large-scale optimization problem, one feasible approach is to use a two-stage process. The first stage pre-selects a small subset of N genes without considering the classifier design. The second stage further reduces the subset while maximizing classification accuracy (Vinterbo et al., 2005; Ho et al., 2006). Ho et al. (2006) proposed an IGA-based method for designing an interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis.

Another existing approach without the pre-selection of genes is to encode *max* integer variables rather than N binary ones into a GA-chromosome where *max* is a pre-specified maximal number of selected genes. GA/MLHD encoded an additional integer variable $G_{\max} \in [1, \max]$ into a GA-chromosome to further reduce the number of genes by only selecting $G_{\max}$ genes corresponding to the first $G_{\max}$ variables in the GA-chromosome (Ooi and Tan, 2003). The objective of GeneSelect aims to select a minimal number of genes while maximizing classification accuracy without using the pre-selection method. We would hybridize the advantages of binary and integer encoding schemes while cooperating with the advantage of IGA at the same time. Let a candidate solution X be encoded into a GA-chromosome, as shown in Fig. 1.

The proposed GA-chromosome encoding scheme consists of both control genes and parametric genes where this gene is a commonly used term of GA, named GA-gene in this paper for discrimination. The max control GA-genes b_i are binary variables where the constant max is pre-defined by design-

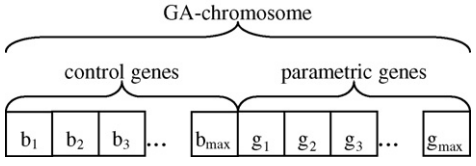

Fig. 1. An illustration of GA-chromosome encoding scheme in GeneSelect.

ers. The parametric GA-genes $g_i \in [1, N]$ are serial numbers of genes in the microarray data. The control GA-gene b_i is used to determine whether the corresponding GA-gene g_i is selected or not. The advantage of using control GA-genes rather than the control variable $G_{\max}$ in GA/MLHD is that each parametric GA-gene has an equal probability to be evaluated that is beneficial to GA, especially to the used IGA.

2.2. Fitness function

Fitness function is the guide of GA's operations to search for optimal solutions. For maximizing classification accuracy $R(\%)$ and minimizing the number G of selected genes, the fitness function f is a weighted sum with a weight w as follows:

$$\max f(X) = R - wG. \tag{1}$$

G is the sum of the values of control GA-gene b_i and R is the accuracy of the combined classifier using these G corresponding genes. The penalty term wG is to further minimize G while maximizing R. For convenient comparisons, we use the same accuracy estimation method and maximum likelihood classification with those of GA/MLHD (Ooi and Tan, 2003). Therefore, the used fitness function using $R = (R_{CV} + R_{IT})/2$ is as follows:

$$\max f(X) = R_{CV} + R_{IT} - wG, \tag{2}$$

where R_{CV} is the correct rate of cross-validation and R_{IT} is the correct rate of pseudo independent test. Note that the fitness function of GA/MLHD is to maximize $R_{CV} + R_{IT}$ only without using the penalty term wG in (2). The accuracy estimations of R_{CV} and R_{IT} can be referred to (Ooi and Tan, 2003) where all the entries in a dataset are split into cross-validation and pseudo independent test sets at a ratio of 2:1 in the gene selection process.

Considering the tradeoff between the accuracy and gene number, the weight w can be adaptively adjusted according to the preference of classifier designers. In this study, to avoid from sacrificing classification accuracy for a small reduction in the number G of selected genes, the fitness function aims to first maximize $R_{CV} + R_{IT}$ where the maximal value is 200 and then to minimize the number G of selected genes. Generally, the number of microarray samples is very small that leads to a great lose of the value of $R_{CV} + R_{IT}$ for one incorrectly classified sample. Therefore, the weight $w = 0.5$ in this study. If the number of samples is significantly increased, the value of w can be set to 0.1 or smaller (Ho et al., 2006).

2.3. IGA for gene selection

The used intelligent genetic algorithm (IGA) is a specific variant of the intelligent evolutionary algorithms (Ho et al., 2004b) for GeneSelect. High performance of IGA mainly arises from an efficient intelligent crossover operation (ICX). ICX is based on orthogonal experimental design (OED) to solve intractable optimization problems comprising lots of design parameters. How to efficiently design a problem-dependent ICX is the major concern in using the intelligent evolutionary algorithm. The proposed ICX with the feasibility maintenance of GA-chromosomes is presented, while the merits of OED and the superiority of ICX can be further referred to (Ho et al., 2004b).

2.3.1. Orthogonal experimental design

An efficient way to study the effects of several factors (variables) simultaneously is to use OED based on orthogonal array (OA) and factor analysis. The two-level OAs used in IGA are concisely described below. Let there be α factors, with two levels each. Generally, to use an OA of α factors, we obtain an integer $M = 2^{\lceil \log_2(\alpha+1) \rceil}$ where the bracket represents an upper ceiling operation, build an OA $L_M(2^{M-1})$ with M rows and $M-1$ columns, use the first α columns, and ignore the other $M-\alpha-1$ columns. In this study, we used $\alpha = 15$. Therefore, $M = 16$ and then the OA $L_{16}(2^{15})$ is used.

After proper tabulation of experimental results, the summarized data are analyzed using factor analysis to determine the relative effects of levels of various factors as follows. Let f_t denote a fitness function value of the combination t, where $t = 1, \ldots, M$. Define the main effect of factor i with level k as S_{ik} where $i = 1, \ldots, \alpha$:

$$S_{ik} = \sum_{t=1}^{M} f_t W_t \tag{3}$$

where $W_t = 1$ if the level of factor i of combination t is k; otherwise, $W_t = 0$. Since the fitness function (2) is to be maximized, level 1 of factor i makes a better contribution to the fitness function than level 2 of factor i does when $S_{i1} > S_{i2}$. If $S_{i1} < S_{i2}$, level 2 is better. If $S_{i1} = S_{i2}$, levels 1 and 2 have the same contribution. The main effect reveals the individual effect of a factor. The most effective factor i has the largest main effect difference $\mathrm{MED}_i = |S_{i1} - S_{i2}|$. After the better one of two levels of each factor is determined, an efficient combination consisting of all factors with the better levels can be easily derived.

2.3.2. Intelligent crossover operation

Like traditional GAs, two parents P_1 and P_2 produce two children C_1 and C_2 in one ICX. Generally, let the OA $L_{n+1}(2^n)$ be used where $n = 2^k - 1$ and k is an integer for efficient use of all columns of OA. One ICX takes $n+2$ fitness evaluations to explore the search space of 2^n combinations. The following steps describe the used ICX:

- Step 1: Move all parametric GA-genes g_i which appear in both P_1 and P_2 to the end of the GA-chromosome. The corresponding control GA-genes b_i are moved to the corresponding positions in the GA-chromosome. These common control and parametric GA-genes in P_1 and P_2 are not participated in this ICX temporally.
- Step 2: Randomly divide all participated control GA-genes and parametric GA-genes into $(n-1)/2$ and $(n+1)/2$ GA-gene segments, respectively. One GA-gene segment is regarded as a factor in OA. Let levels 1 and 2 of factor j

represent the jth GA-gene segments coming from P_1 and P_2, respectively.

- Step 3: Evaluate the fitness value f_t of the combinations corresponding to the experiments t, where $t = 2, \ldots, n+1$. The value f_1 is the fitness value of P_1.
- Step 4: Compute the main effect S_{ik} where $i = 1, \ldots, n$ and $k = 1, 2$.
- Step 5: Determine the better one of two levels of each factor.
- Step 6: The GA-chromosome of C_1 is formed using the combination of the better GA-genes from the derived corresponding parents.
- Step 7: The GA-chromosome of C_2 is formed similarly as C_1, except that the factor with the smallest main effect difference adopts the other level.
- Step 8: The best two individuals among P_1, P_2, C_1, C_2, and n combinations of OA are selected for elitist strategy. The final children C_1 and C_2 are the selected individuals where the control and parametric GA-genes not participated in Step 1 are appended.

In Step 1, since all parametric GA-genes participated in ICX are different in two parents, it is impossible to produce infeasible children having duplicated GA-genes. Therefore, Step 1 can ensure that all the recombined GA-chromosomes are always feasible without using the repair operation or penalty approach, which makes ICX more efficient. Note that the recombination of parents in GA/MLHD would result in infeasible children. Therefore, an additional repair operation is needed in GA/MLHD.

2.3.3. Intelligent genetic algorithm

The simple IGA used is given as follows:

- Step 1: Randomly generate an initial population with N_{pop} individuals.
- Step 2: Evaluate fitness values of all individuals.
- Step 3: Binary tournament selection without replacement is adopted.
- Step 4: Randomly select $P_c N_{pop}$ individuals where P_c is a crossover probability. Perform ICXs for all selected pairs of parents.
- Step 5: Apply a conventional bit-inverse mutation operator to the population using a mutation probability P_m. To prevent the best fitness value from deteriorating, mutation is not applied to the best individual.
- Step 6: Termination test: if a pre-specified termination condition is satisfied, stop the algorithm. Otherwise, go to step 2.

3. Experiments

3.1. Evaluation of GeneSelect

There are four major differences between the two methods IGA/MLHD and GA/MLHD, listed in Table 1.

Table 1
Major differences between GA/MLHD and IGA/MLHD

Differences	GA/MLHD	IGA/MLHD
GA-chromosome encoding	Control variable	Control gene
Fitness function	$R_{CV} + R_{IT}$	$R_{CV} + R_{IT} - wG$
GA crossover operation	UX	ICX
Feasibility maintenance	Repair	Reordering

The four improvements are the merits of GeneSelect. In the following simulations in this subsection, we evaluate each of the four refinements by using the same GA/MLHD classifier that only one refinement is compared at a time, unless otherwise specified. The GA/MLHD classifier is obtained by running the program provided by Ooi and Tan (2003). All the experimental results are the averaged values of 30 independent runs. The typical results using the dataset NCI60 (Ross et al., 2000) with nine classes are given for illustrating the performance comparisons.

3.1.1. GA-chromosome encoding

The performance comparison is obtained by running the GA/MLHD method with two GA-chromosome encoding methods: the control variable G_{max} (Ooi and Tan, 2003) and control gene (GeneSelect). Experimental results of the two encoding methods are shown in Fig. 2. The results reveal that the control gene method of GeneSelect has better performance.

Because the GA-genes in the control gene method increase the length of GA-chromosome, it would result in a larger search space. Since IGA with ICX using a

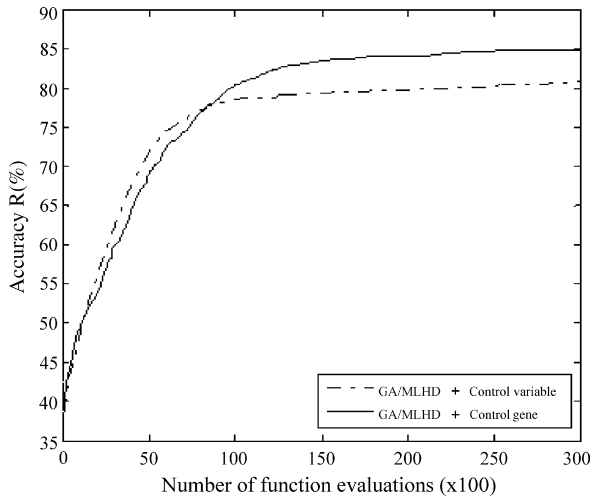

Fig. 2. Performance comparison between the control variable and control gene methods.

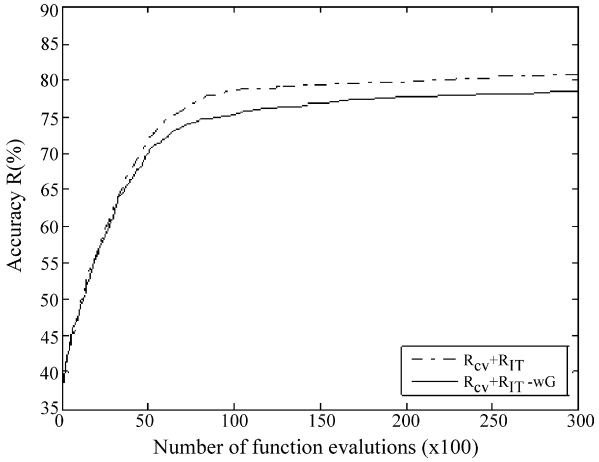

Fig. 3. Performance evaluation of using the simple GA with and without a penalty term.

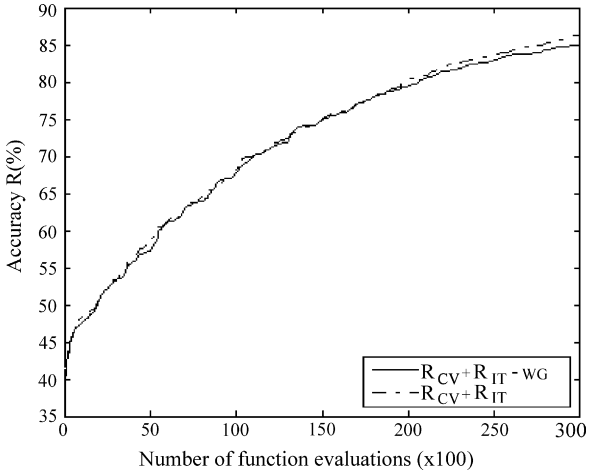

Fig. 4. Performance evaluation of using IGA with and without a penalty term wG.

divide-and-conquer approach can efficiently solve large parameter optimization problems, IGA would make the control gene method more efficient than the simple GA of GA/MLHD, especially when the value of max is large.

3.1.2. Fitness function

The fitness function has an additional penalty term wG which can further minimize the number G of selected genes provided that the used GA is efficient enough to maximize the accuracy. Fig. 3 shows the performance comparison of GA/MLHD with and without using the penalty term. In Fig. 3, GA/MLHD has the mean accuracy 81.2% with the number 17.09 of selected genes. If the penalty term is used, the results of accuracy and gene number are 78.2 and 13.05%, respectively. The number of genes becomes smaller but the accuracy is also reduced. This scenario results from that the search efforts of GA aim to achieve the two conflicting objectives simultaneously. However, the performance of using the penalty term can be improved by using a more efficient GA, such as IGA.

Instead of the simple GA, the performance comparison of using IGA with and without the penalty term in two classifiers is shown in Fig. 4. The performance of using the penalty term is only slightly worse than that of using no penalty term. The result shows that IGA can minimize the number of selected genes using the fitness function with the penalty term while maximizing the accuracy. The improved performance of IGA with the penalty term is 85.0% with 12.95 genes. In other words, IGA instead of the simple GA can make the fitness function with the penalty term more efficient.

3.1.3. Crossover operation

High performance of the used IGA in GeneSelect arises from the intelligent crossover operation (ICX) based on OAs and factor analysis. For observing the effectiveness of ICX only, the performances are obtained using GA/MLHD with the two crossover operations: uniform crossover (UX) and ICX, shown in Fig. 5. Because ICX needs extra fitness evaluations per crossover operation, the number of generations using GA with ICX is less than GA with UX using the same number of fitness evaluations. Therefore, UX performs well in the early evolution process but ICX performs well in the later evolution process. If the computation time is sufficient, ICX is much better than UX to obtain a satisfactory solution.

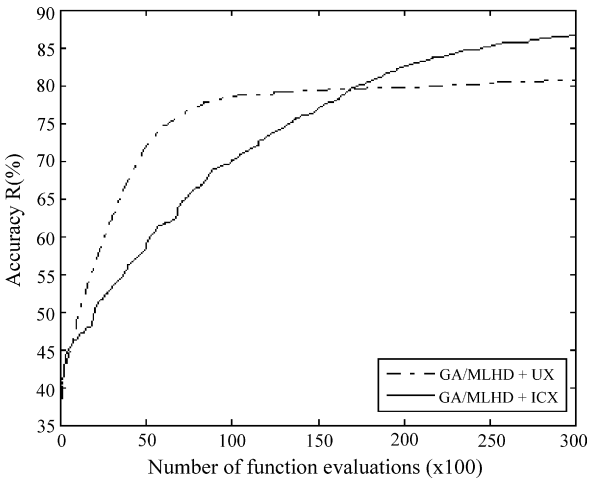

Fig. 5. Performance comparison between uniform crossover (UX) and intelligent crossover operation (ICX) using GA/MLHD.

3.1.4. Feasibility maintenance

Feasibility maintenance is to make sure that no two parametric GA-genes g_i and g_j in a GA-chromosome are the same. Obviously, random recombination in the crossover step of GA often makes the children infeasible. Generally, it is commonly used to repair the infeasible GA-chromosomes by replacing the duplicated GA-gene with another one like GA/MLHD. The proposed GA-chromosome reordering method can make the recombination of GA-chromosomes always feasible. Because the genetic search is always confined in a feasible search space, the performance of GA would be higher by focusing on searching for optimal solutions. The performance comparison of using the repair and reordering methods is shown in Fig. 6. It is obvious that the proposed GA-chromosome reordering method is effective.

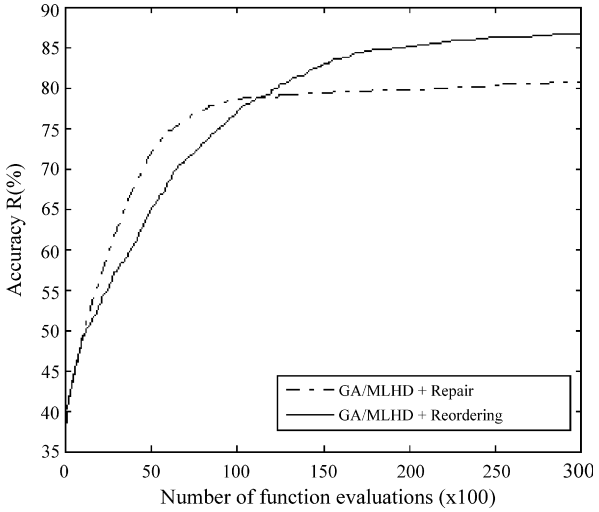

Fig. 6. Performance comparison between the repair and reordering methods using GA/MLHD.

3.2. Performance comparisons with GA/MLHD

The proposed IGA-based gene selection method (GeneSelect) is compared with the GA-based one using the same MLHD classifier and estimation of classification accuracy (Ooi and Tan, 2003). The parameter settings of IGA are as follows: $N_{pop} = 20$, $P_c = 1.0$, $P_m = 0.002$, max = 50, and $L_{16}(2^{15})$ with $n = 15$. Two compared GAs used the same number of fitness evaluations (Ooi and Tan, 2003) as the stopping condition for fair comparisons. The 11 cancer-related human gene expression datasets gleaned from the literature are described in Table 2. Each of the two classifiers GA/MLHD and IGA/MLHD performed 30 independent runs.

Fig. 7 shows the average convergence performance of two methods on NCI60. It is obvious that IGA/MLHD is significantly better than GA/MLHD where the accuracy is up to 90%. From Figs. 2–7, it can be found that IGA/MLHD using the four improvements simultaneously is better than GA/MLHD using only one of four improvements. In order to evaluate the robustness of the IGA-based approach on selected genes, we calculated the frequency of selected genes from solutions of 100 independent runs on NCI60, shown in Fig. 8. This analysis reveals that the frequency of relevant genes selected by IGA/MLHD is higher than GA/MLHD. Because the number of tissue samples is usually much smaller than

Table 2
Cancer-related human gene expression datasets

Dataset name	Diagnostic task	Number of samples	Number of genes	Number of classes	Reference
NCI60	Nine various human tumor types	60	5725	9	Ross et al. (2000)
14_Tumors	Fourteen various human tumor types and 12 normal tissue types	308	15009	26	Ramaswamy et al. (2001)
11_Tumors	Eleven various human tumor types	174	12533	11	Su et al. (2001)
Brain_Tumor1	Five human brain tumor types	90	5920	5	Pomeroy et al. (2002)
Brain_Tumor2	Four malignant glioma types	50	10367	4	Nutt et al. (2003)
Leukemia1	Acute myelogenous leukemia (AML), Acute lymphoblastic leukemia (ALL) B cell, and ALL T-cell	72	5327	3	Golub et al. (1999)
Leukemia2	AML, ALL, and mixed-lineage leukemia (MLL)	72	11225	3	Armstrong et al. (2002)
Lung_Cancer	Four lung cancer types and normal tissues	203	12600	5	Bhattacharjee et al. (2001)
SRBCT	Small, round blue cell tumors of childhood	83	2308	4	Khan et al. (2001)
Prostate_Tumor	Prostate tumor and normal tissue	102	10509	2	Singh et al. (2002)
DLBCL	Diffuse large B cell lymphomas and follicular lymphomas	77	5469	2	Shipp et al. (2002)

Table 3
Experiment results of GA/MLHD and IGA/MLHD where the accuracy $R=(R_{CV}+R_{IT})/2$ and G is the number of selected genes

Dataset	GA/MLHD						IGA/MLHD					
	Best		Mean		Variance		Best		Mean		Variance	
	R (%)	G	R (%)	G	R (%)	G	R (%)	G	R (%)	G	R (%)	G
NCI60	82.15	17	72.8	16.3	11.37	16.7	96.1	12	91.75	12.8	7.12	12.3
14_Tumors	62.6	35	54.25	35.2	31.32	30.2	82.7	29	74.25	27.4	16.95	42.8
11_Tumors	88.15	32	81.45	31.2	19.87	22.6	100	28	94.85	28.3	6.45	18.5
Brain_Tumor1	96.7	14	93.55	11.3	3.85	11.4	100	6	99.2	6.7	2.37	6.7
Brain_Tumor2	100	7	97.1	10.7	1.90	7.9	100	4	99.65	4.5	1.90	4.4
Leukemia1	100	9	97.85	12.7	1.15	8.2	100	2	100	3.8	0.85	3.2
Leukemia2	100	8	99.2	13.6	0.60	4.1	100	3	100	3.1	0.42	1.2
Lung_Cancer	97.1	10	94.15	12.1	3.82	14.4	100	6	99.45	6.9	1.32	3.6
SRBCT	100	11	98.15	9.8	2.92	6.6	100	5	100	5.8	1.52	2.3
Prostate_Tumor	96.4	5	92.6	7.8	2.22	3.5	100	3	99.4	4.4	1.80	0.7
DLBCL	100	6	97.95	6.3	0.80	0.8	100	2	99.7	4.8	0.35	0.3
Mean	93.00	14.00	89.00	15.18	7.25	11.49	98.07	9.09	96.20	9.86	3.73	8.72

The best performance has the largest value of R.

the number of genes, it may occur that there are multiple different sets with the same small number of genes having the same high accuracy. The frequency of selected genes can be further improved if the number of samples is increased or the system uncertainty is taken into account such as by averaging over multiple sets of potentially overlapping relevant genes (Yeung et al., 2005).

The results on the 11 datasets are given in Table 3 where the accuracy $R=(R_{CV}+R_{IT})/2$ and G is the number of selected genes. The experimental results show that IGA/MLHD is superior to GA/MLHD in terms of the number of selected genes (9.86 versus 15.18 on average), classification accuracy (96.20% versus 89.00% on average), and robustness of selected genes and accuracy. For the best performance in terms of R, IGA/MLHD has $R=98.07\%$ with $G=9.09$ genes averagely. For the datasets with a small number (≤ 5) of classes, IGA/MLHD can obtain 100% accuracy with a very small number of genes (≤ 6).

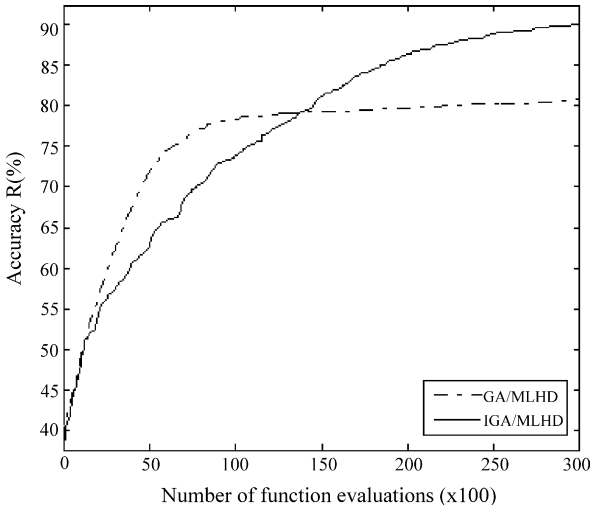

Fig. 7. Average convergence performance of two methods GA/MLHD and IGA/MLHD on NCI60.

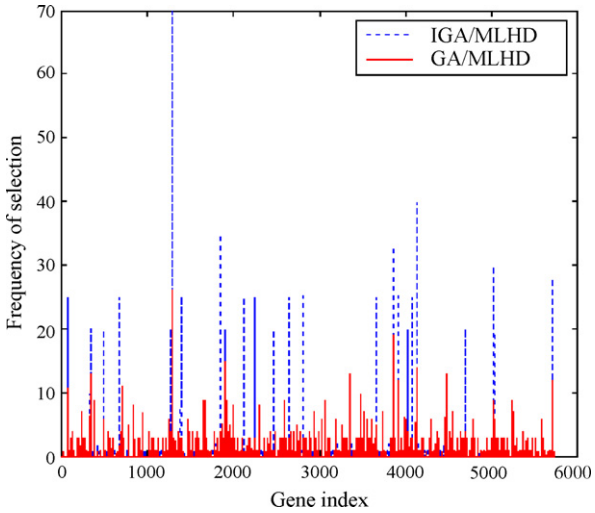

Fig. 8. Frequency (%) of selected genes from solutions of 100 independent runs on NCI60.

4. Conclusions

We have proposed a generalized gene selection method named GeneSelect from microarray data which can be combined with the optimal designs of various

multiclass classification methods. The merits of GeneSelect are threefold: an efficient encoding scheme of GA-chromosomes, an associated fitness function with a penalty term, and a powerful intelligent genetic algorithm (IGA) with the intelligent crossover operation having the ability of maintaining feasibility without using repair operations. In this study, GeneSelect is applied to the maximum likelihood classification (MLHD) for comparisons with an existing method (GA/MLHD). The proposed IGA/MLHD method can simultaneously optimize gene selection and tissue classification for microarray data analyses. After computer simulation using 11 benchmark datasets, it reveals that IGA/MLHD could obtain not only higher classification accuracy but also a smaller number of relevant genes than the existing method GA/MLHD. In addition, the IGA-based gene selection method is more robust in both the frequency of selected relevant genes as well as the classification accuracy. Therefore, GeneSelect is an efficient method in designing classifiers for analyses of microarray data.

Acknowledgements

The authors would like to thank Ooi and Tan for providing the program of their method GA/MLHD.

References

Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., Korsmeyer, S.J., 2002. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat. Genet. 30 (1), 41–47.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J., Meyerson, M., 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. U.S.A. 98 (24), 13790–13795.

Goldberg, D.E., 1989. Genetic Algorithms in search. In: Optimization and Machine Learning. Addison-Wesley Publishing Company.

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

Ho, S.-Y., Liu, C.-C., Liu, S., 2002. Design of an optimal nearest neighbor classifier using an intelligent genetic algorithm. Pattern Recogn. Lett. 23, 1495–1503.

Ho, S.-Y., Chen, J.-H., Huang, M.-H., 2004a. Inheritable genetic algorithm for bi-objective 0/1 combinatorial optimization problems and its applications. IEEE Trans. Syst. Man Cybern. B 34 (1), 609–620.

Ho, S.-Y., Shu, L.-S., Chen, J.-H., 2004b. Intelligent evolutionary algorithms for large parameter optimization problems. IEEE Trans. Evol. Comput. 8 (6), 522–541.

Ho, S.-Y., Hsieh, C.-H., Chen, H.-M., Huang, H.-L. (in press). Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. BioSystem.

Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7 (6), 658–659.

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, L., Weinberg, C.R., Darden, T.A., Pedersen, L.G., 2001. 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.

Li, T., Zhang, C., Ogihara, M., 2004. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20, 2429–2437.

Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Cairncross, J.G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M.E., Batchelor, T.T., Black, P.M., Deimling, A.V., Pomeroy, S.L., Golub, T.R., Louis, D.N., 2003. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res. 63 (7), 1602–1607.

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

Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., Golub, T.R., 2002. Prediction of central nervous system embryonal tumor outcome based on gene expression. Nature 415 (6870), 436–442.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., Golub, T.R., 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl Acad. Sci. U.S.A. 98 (26), 15149–15154.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffery, S.S., Van de Rijn, M., Waltham, M., 2000. Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet. 24, 227–235.

Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., Golub, T.R., 2002. Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. Nat. Med. 8 (1), 68–74.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T., Sellers, W., 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209.

Su, A.I., Welsh, J.B., Sapinoso, L.M., Kern, S.G., Dimitrov, P., Lapp, H., Schultz, P.G., Powell, S.M., Moskaluk, C.A., Frierson Jr., H.F., Hampton, G.M., 2001. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res. 61 (20), 7388–7393.

Vinterbo, S.A., Kim, E.Y., Ohno-Machado, L., 2005. Small, fuzzy and interpretable gene expression based classifiers. Bioinformatics 21, 1964–1970.

Yeung, K.Y., Bumgarner, R.E., 2003. Multiclass classification of microarray data with repeated measurements: application to cancer. Genome Biol 4, R83.

Yeung, K.Y., Bumgarner, R.E., Raftery, A.E., 2005. Bayesian model averaging: development of an improved multiclass, gene selection and classification tool for microarray data. Bioinformatics 21 (10), 2394–2402.