Available online at www.sciencedirect.com
ScienceDirect
BioSystems 85 (2006) 165–176
BioSystems
www.elsevier.com/locate/biosystems

Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis

Shinn-Ying Ho [a,b,*], Chih-Hung Hsieh [b], Hung-Ming Chen [c], Hui-Ling Huang [d]

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

Received 10 August 2005; received in revised form 14 December 2005; accepted 3 January 2006

Abstract

An accurate classifier with linguistic interpretability using a small number of relevant genes is beneficial to microarray data analysis and development of inexpensive diagnostic tests. Several frequently used techniques for designing classifiers of microarray data, such as support vector machine, neural networks, *k*-nearest neighbor, and logistic regression model, suffer from low interpretabilities. This paper proposes an interpretable gene expression classifier (named iGEC) with an accurate and compact fuzzy rule base for microarray data analysis. The design of iGEC has three objectives to be simultaneously optimized: maximal classification accuracy, minimal number of rules, and minimal number of used genes. An "intelligent" genetic algorithm IGA is used to efficiently solve the design problem with a large number of tuning parameters. The performance of iGEC is evaluated using eight commonly-used data sets. It is shown that iGEC has an accurate, concise, and interpretable rule base (1.1 rules per class) on average in terms of test classification accuracy (87.9%), rule number (3.9), and used gene number (5.0). Moreover, iGEC not only has better performance than the existing fuzzy rule-based classifier in terms of the above-mentioned objectives, but also is more accurate than some existing non-rule-based classifiers.

Keywords: Fuzzy classifier; Gene expression; Intelligent genetic algorithm; Microarray data analysis; Pattern recognition

1. Introduction

Microarray is a useful technique for measuring expression data of thousands of genes simultaneously. Microarray gene expression profiling technology is one of the most important research topics in clinical diagnosis of disease. Gene expression data provide valuable information in the understanding of genes, biological networks, and cellular states. One goal in analyzing expression data is to determine how the expression of any particular gene might affect the expression of other genes in the same genetic network (Ressom et al., 2003; Woolf and Wang, 2000; Kauffman et al., 2003; Wahde and Hertz, 2000). Another goal is to determine how genes are expressed as a result of certain cellular conditions (e.g., how genes are expressed in diseased and healthy cells) (Creighton and Hanash, 2003).

The practical applications of microarray gene expression profiles include management of cancer and infectious diseases. The prediction of the diagnostic category of a tissue sample from its expression array phenotype from tissues in identified categories is known as classification. Because the number of tissue samples is usually much smaller than the number of genes, it may occur

* 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.2006.01.002

that there are multiple different sets with the same small number of genes having the same high classification accuracy (Yeung et al., 2005). In analyzing expression data by designing classifiers, it is better to provide additional biological knowledge associated for verifying the selected genes rather than the emphasis of both high classification accuracy and small number of used genes only. In this study, designing an accurate and compact fuzzy rule-based classifier with linguistic interpretability using a small number of relevant genes is investigated, which is beneficial to microarray data analysis and development of inexpensive diagnostic tests. The desirable classifier is a set of fuzzy rules with linguistic interpretability where each rule is as the similar form: if gene *A* is up-regulated and gene *B* is down-regulated, then the probability of disease *X* is high.

Merz (2003) applied memetic algorithms to the minimum sum-of-squares clustering problem for gene expression profile analysis. Recently, some supervised machine learning techniques, such as support vector machine (SVM), neural networks (NN), *k*-nearest neighbor (*k*-NN), and logistic regression have been used in designing gene expression data classifiers (Statnikov et al., 2005; Vinterbo et al., 2005). Liu et al. (2004) proposed a feature selection method which combines top-ranked, test-statistic, and principle component analysis in conjunction with ensemble NN to design classifiers. Zhou and Mao (2005) suggested a filter-like evaluation criterion, called LS Bound measure, derived from leave-one-out procedure of least squares support vector machines (LS-SVMs), which provides gene subsets leading to more accurate classification. Liu et al. (2005a) combined the entropy-based feature selection method using simulated annealing and *k*-NN classifier for cancer classification. Liu et al. (2005b) proposed a hybrid method which combines GA and SVM for multi-class cancer categorization.

Statnikov et al. (2005) investigated some existing multi-class classification methods and indicated that the multi-category SVM is the most effective classifier for tumor classification in terms of classification accuracy using a very large number of genes. However, given thousands of genes, only a small number of genes show strong correlation with a certain phenotype (Ding, 2003). To advance the classification performance using a small number of genes, it is better to take both gene selection and classifier design into account simultaneously (Deb and Reddy, 2003). Li et al. (2001) proposed a hybrid method of the genetic algorithm (GA)-based gene selection and *k*-NN classifier to assess the importance of genes for classification. Ooi and Tan (2003) proposed a maximal likelihood based method for the multi-category prediction of gene expression data.

However, learning results of the above-mentioned classifiers containing equations involving several coefficients, interaction terms, and constants cannot be summarized into linguistically interpretable forms for biologists and biomedical scientists (Vinterbo et al., 2005). Li et al. (2003) used a tree structure to classify microarray samples. Hvidsten et al. (2003) proposed learning rule-based models of biological process from gene expression time profiles using gene ontology. Vinterbo et al. (2005) presented a rule-induction and filtering strategy to obtain an accurate, small, and interpretable fuzzy classifier using a grid partition of feature space, compared with the classifier of logistic regression.

In this paper, we propose an interpretable gene expression classifier (named iGEC) with an accurate and compact fuzzy rule base using a scatter partition of feature space for microarray data analysis. Because gene expression data have the property of natural clustering, fuzzy classifiers using a scatter partition of feature spaces often have a smaller number of rules than those using grid partition (Ho et al., 2004a). The design of iGEC has three objectives to be simultaneously optimized: maximal classification accuracy, minimal number of rules, and minimal number of used genes. In designing iGEC, the flexible membership function, fuzzy rule, and gene selection are simultaneously optimized. An "intelligent" genetic algorithm IGA is used to efficiently solve the design problem with a large number of tuning parameters (Ho et al., 2004a).

The performance of iGEC is evaluated using eight gene expression data sets. It is shown that iGEC has an accurate, concise, and interpretable rule base (1.1 rules per class) averagely in terms of test classification accuracy (87.9%), rule number (3.9), and used gene number (5.0). Moreover, iGEC not only has better performance than the classifier (Vinterbo et al., 2005) in terms of the above-mentioned objectives, but also is more accurate than some existing non-rule-based classifiers (*k*-NN and NNs).

2. Methods

High performance of iGEC mainly arises from two aspects. One is to simultaneously optimize all parameters in the design of iGEC where all the elements of the fuzzy classifier design have been moved in parameters of a large parameter optimization problem. The other is to use an efficient optimization algorithm IGA which is a specific variant of the intelligent evolutionary algorithm (Ho et al., 2004b). The intelligent evolutionary algorithm uses a divide-and-conquer strategy to effectively solve large parameter optimization problems. IGA

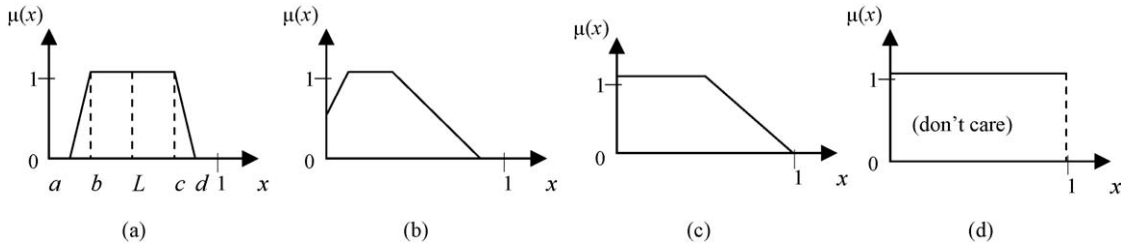

Fig. 1. Illuminations of FGPMF: (a) $a > 0$ and $d < 1$; (b) $a < 0 < b$; (c) $b \leq 0$; (d) $b \leq 0$ and $c \geq 1$.

is shown to be effective in the design of accurate classifiers with a compact fuzzy-rule base using an evolutionary scatter partition of feature space (Ho et al., 2004a).

2.1. Flexible membership function

The classifier design of iGEC uses flexible generic parameterized fuzzy regions which can be determined by flexible generic parameterized membership functions (FGPMFs) and a hyperbox-type fuzzy partition of feature space. Each fuzzy region corresponds to a parameterized fuzzy rule. In this study, each value of gene expression is normalized into a real number in the unit interval [0,1]. An FGPMF with a single fuzzy set is defined as

$$\mu(x) = \begin{cases} 0 & \text{if } x \leq a \text{ or } x \geq d \\ \dfrac{x-a}{b-a} & \text{if } a < x < b \\ \dfrac{d-x}{d-c} & \text{if } c < x < d \\ 1 & \text{if } b \leq x \leq c \end{cases} \quad (1)$$

where $x \in [0, 1]$ and $a \leq b \leq c \leq d$. The variables a, b, c, and d determining the shape of a trapezoidal fuzzy set are the parameters to be optimized. It is well recognized that confining evolutionary searches within feasible regions is often much more reliable than penalty approaches for handling constrained problems (Michalewicz et al., 1996). Therefore, five parameters $V^1, V^2, \ldots, V^5 \in [0, 1]$ without constraints instead of a, b, c, and d are encoded into a GA-chromosome for facilitating IGA. Let an additional variable $L = V^1$ which determines the location of the fuzzy set characterizing the occurrence of training patterns. When V^i are obtained, variables a, b, c, and d can be derived as follows: $a = L - (V^2 + V^3)$, $b = L - V^3$, $c = L + V^4$, and $d = L + (V^4 + V^5)$. This transformation can always make the derived values of a, b, c, and d feasible and reduce interactions among encoded parameters of GA-chromosomes. Some illuminations of FGPMF are shown in Fig. 1.

2.2. Fuzzy rule and fuzzy reasoning method

The following fuzzy if–then rules for n-dimensional pattern classification problems are used in the design of iGEC:

$$R_j : \text{If } x_1 \text{ is } A_{j1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{jn} \text{ then class } CL_j \text{ with } CF_j, \quad j = 1, \ldots, N.$$

where R_j is a rule label, x_i denotes a gene variable, A_{ji} is an antecedent fuzzy set, C is a number of classes, $CL_j \in \{1, \ldots, C\}$ denotes a consequent class label, CF_j is a certainty grade of this rule in the unit interval [0, 1], and N is a number of initial fuzzy rules in the training phase.

To enhance interpretability of fuzzy rules, linguistic variables in fuzzy rules can be used. Each variable x_i has a linguistic set $U = \{\text{L, ML, M, MH, H}\}$. Each linguistic value of x_i equally represents 1/5 of the domain [0, 1]. Following the quantization criterion, we can consider genes to be regulated according to a qualitative level. For example, x_i is Low for down-regulated genes; x_i is Medium for neutral genes; and x_i is High for up-regulated genes. An antecedent fuzzy set $A_{ji} \in A_u$ where A_u denotes a set of subsets of U. Examples of linguistic antecedent fuzzy sets are shown in Fig. 2.

In the training phase, all the variables CL_j and CF_j are treated as parametric genes of GA (GA-genes) encoded in chromosomes of GA (GA-chromosomes) and their values are

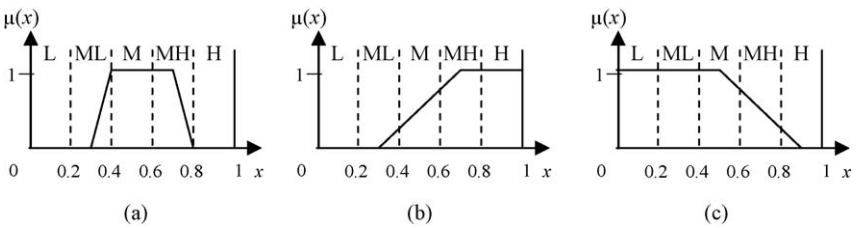

Fig. 2. Examples of an antecedent fuzzy set A_{ji} with linguistic values (L: low, ML: medium low, M: medium, MH: medium high, H: high): (a) A_{ji} represents {ML, M, MH}; (b) A_{ji} represents {ML, M, MH, H}, i.e., not Low; (c) A_{ji} represents {L, ML, M, MH, H} or ALL.

obtained using IGA. The following fuzzy reasoning method is adopted to determine the class of an input pattern $x_p = (x_{p1}, x_{p2}, \ldots, x_{pn})$ based on voting using multiple fuzzy if–then rules:

Step 1: Calculate score $S_{\text{Class } v}(v = 1, \ldots, C)$ for each class as follows:

$$S_{\text{Class } v} = \sum_{\substack{R_j \in FC \\ CL_j = \text{Class } v}} \mu_j(x_p) CF_j,$$

$$\mu_j(x_p) = \prod_{i=1}^{n} \mu_{ji}(x_{pi}), \tag{2}$$

where FC denotes the fuzzy classifier, the scalar value and $\mu_{ji}(\cdot)$ represents the membership function of the antecedent fuzzy set A_{ji}.

Step 2: Classify x_p as the class with a maximal value of $S_{\text{Class } v}$.

2.3. Fitness function and GA-chromosome representation

We define the fitness function $Fit()$ of IGA for designing iGEC as follows:

$$\max \; Fit(FC) = NCP - W_r N_r - W_f N_f \tag{3}$$

where W_r and W_f are positive weights. In this study, the fitness function is used to optimize the three objectives in the following order: to maximize the number NCP of correctly classified training patterns, to minimize the number N_r of fuzzy rules, and to minimize the number N_f of selected genes. Generally, the final number of fuzzy rules is smaller than 10. Therefore, we set $W_r = 0.1$ to ensure that classification accuracy has the first priority to be optimized. When the two objectives NCP and N_r are simultaneously optimized for microarray data, the best number of used genes is almost determined. Hence, a very small value 0.001 is set to W_f. The sensitive analysis about the different settings of W_r and W_f can be referred to Ho et al. (2004a). It is shown that various combinations of feasible settings of W_r and W_f are not sensitive to performance of the fuzzy-rule based classifier (Ho et al., 2004a).

A GA-chromosome consists of control GA-genes for selecting useful genes and significant fuzzy rules, and parametric GA-genes for encoding the membership functions and fuzzy rules. The control GA-genes comprise two types of parameters. One is parameter r_j, $j = 1, \ldots, N$, represented by one bit for eliminating unnecessary fuzzy rules. If $r_j = 0$, the fuzzy rule R_j is excluded from the rule base. Otherwise, R_j is included. The other is parameter f_i, $i = 1, \ldots, n$, represented by one bit for eliminating useless genes. If $f_i = 0$, the gene x_i is excluded from the classifier. Otherwise, x_i is included. The parametric GA-genes consist of three types: $V_{ji}^k \in [0, 1]$, $k = 1, \ldots, 5$, for determining the antecedent fuzzy set A_{ji}; CL_j for determining the consequent class label of rule R_j; and $CF_j \in [0, 1]$ for determining the certainty grade of rule R_j; where $j = 1, \ldots, N$ and $i = 1, \ldots, n$. A rule base with N fuzzy rules is represented as an individual, as shown in Fig. 3. The number of encoding parameters to be optimized is equal to $N_p = n + 3N + 5Nn$. A GA-chromosome representation uses a binary string for encoding control and parametric GA-genes. There are eight bits for encoding one of parameters V_{ji}^k and CF_j. Since each fuzzy region defines a fuzzy rule, the initial setting of N is independent of n but dependent on the number of fuzzy regions. Generally, N is set to the maximal number of possible fuzzy regions. In this study, $N = 3C$. The design of an efficient fuzzy classifier is formulated as a large parameter optimization problem. Once the solution of IGA is obtained, an accurate classifier with a compact fuzzy rule base can be derived.

2.4. IGA for designing iGEC

The main difference between IGA and the traditional GA (Goldberg, 1989) is an efficient intelligent crossover operation. The intelligent crossover is based on orthogonal experimental design to solve intractable optimization problems comprising lots of system parameters. The intelligent crossover is presented while the merits of orthogonal experimental design and the superiority of intelligent crossover can be further referred to Ho et al. (2004a, 2004b).

2.4.1. Orthogonal experimental design

The two-level orthogonal arrays (OAs) used in IGA are described below. Let there be α factors, with two levels each. The total number of level combinations is 2^α for a complete factorial experiment. 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. OA can reduce the number of level combinations for factor analysis. The number of OA combinations required to analyze all individual factors is only $M = O(\alpha)$, where $\alpha + 1 \le M \le 2\alpha$.

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 y_t denote a objective function value of the combination t, where

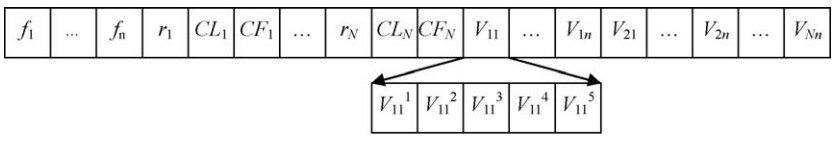

Fig. 3. GA-chromosome representation.

$t = 1, \ldots, M$. Define the main effect of factor i with level k as S_{ik} where $i = 1, \ldots, \alpha$:

$$S_{ik} = \sum_{t=1}^{M} y_t W_t \quad (4)$$

where $W_t = 1$ if the level of factor i of combination t is k; otherwise, $W_t = 0$. Since the fitness function 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 $MED_i = |S_{i1} - S_{i2}|$. 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.

2.4.2. Intelligent crossover

All parameters are encoded into a GA-chromosome using binary codes. Like traditional GAs, two parents P_1 and P_2 produce two children C_1 and C_2 in one crossover operation. Let all encoded parameters be randomly assigned into α groups where each group is treated as a factor. The following steps describe the intelligent crossover operation.

Step 1: Use the first α columns of an OA $L_M(2^{M-1})$.
Step 2: Let levels 1 and 2 of factor i represent the ith groups of parameters coming from parents P_1 and P_2, respectively.
Step 3: Evaluate the fitness values y_t for experiment t where $t = 2, \ldots, M$. The value y_1 is the fitness value of P_1.
Step 4: Compute the main effect S_{ik} where $i = 1, \ldots, \alpha$ 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 $M - 1$ combinations of OA are used as the final children C_1 and C_2 for elitist strategy.

One intelligent crossover operation takes $M + 1$ fitness evaluations, where $\alpha + 1 \leq M \leq 2\alpha$, to explore the search space of 2^α combinations.

2.4.3. Intelligent genetic algorithm

The used IGA is given as follows:

Step 1: Randomly generate an initial population with N_{pop} individuals.
Step 2: Evaluate fitness values of all individuals in the population. Let I_{best} be the best individual in the population.
Step 3: Use the simple ranking selection that replaces the worst $P_s N_{pop}$ individuals with the best $P_s N_{pop}$ individuals to form a new population, where P_s is a selection probability.
Step 4: Randomly select $P_c N_{pop}$ individuals including I_{best}, where P_c is a crossover probability. Perform intelligent crossover operations 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. Implementation and data sets

The parameter settings of IGA from Ho et al. (2004a) are $N_{pop} = 20$, $P_c = 0.7$, $P_s = 1 - P_c$, $P_m = 0.01$, and $\alpha = 15$. Because the search space of optimal design of iGEC is proportional to the number N_p of parameters to be optimized, the stopping condition is suggested to use a fixed number $100N_p$ of fitness evaluations (Ho et al., 2004a) for the following two reasons: (1) for future comparisons with other methods based on the same computation cost; and (2) satisfactory solutions can be obtained which are not sensitive to the number of evaluations used. Of course, if the number of evaluations is increased, the results may be slightly improved. Because of the non-deterministic characteristic of GA, all the experimental results are the average values of 30 independent runs. For each run, a 10-fold cross validation (10-CV) is adopted. Note that the algorithm proposed by Vinterbo et al. (2005) is deterministic that the results are the same for all independent runs.

For comparison, we adopted the same Wilcoxon rank sum test with Vinterbo et al. (2005) as a non-parametric feature pre-selection method. In this study, we pre-selected $n = 10$, 15, 20, and 100 representative genes to evaluate the performance of iGEC. Considering the test accuracy as well as the numbers of rules and genes, $n = 15$ (slightly better) is suggested as the default setting of iGEC in this study. If the number C of classes is further increased (e.g., $C > 10$), the number n is suggested to be proportionally increased.

Table 1 shows the eight data sets from Statnikov et al. (2005), which are available from http://www.gems-systems.org/. The following experiments are designed to evaluate the proposed method using comparisons with some existing rule and non-rule based classifiers. The first comparison is made between iGEC and the Vinterbo's fuzzy rule-based classifier and the second

Table 1
The eight data sets from Statnikov et al. (2005)

No.	Data set	Descriptions	No. of classes	No. of samples	No. of genes	N_p	Reference
1	Brain tumor1	5 human brain tumor types	5	90	5920	1185	Pomeroy et al. (2002)
2	Brain tumor2	4 malignant glioma types	4	50	10367	951	Nutt et al. (2003)
3	DLBCL	Diffuse large b-cell lymphomas and follicular lymphomas	2	77	5469	483	Shipp et al. (2002)
4	Leukemia1	Acute myelogenous leukemia (AML), Acute lympboblastic leukemia (ALL) B-cell, and ALL T-cell	3	72	5327	717	Golub et al. (1999)
5	Leukemia2	AML, ALL, and mixed-lineage leukemia (MLL)	3	72	11225	717	Armstrong et al. (2002)
6	Lung cancer	4 lung cancer types and normal tissues	5	203	12600	1185	Bhattacharjee et al. (2001)
7	Prostate tumor	Prostate tumor and normal tissue	2	102	10509	483	Singh et al. (2002)
8	SRBCT	Small, round blue cell tumors of childhood	4	83	2308	951	Khan et al. (2001)

one between iGEC and the non-rule-based classifiers in Statnikov et al. (2005).

3.2. Results

For comparisons, we conducted two evaluations on the Vinterbo's method using different numbers of pre-selected genes. One is to use 200 pre-selected genes (V200), which is the same with that in Vinterbo et al. (2005). The other is to use 15 genes (V15), which is the same with that of the proposed method. Table 2 shows the statistical results (mean and standard deviation) of iGEC and the Vinterbo's classifier in terms of training accuracy, test accuracy, number of rules, number of genes, and rule number per class. The results of the Vinterbo's classifier were obtained by running the same program provided by Vinterbo et al. (2005). The same data which have the same partition are used for iGEC, V200, and V15. Fig. 4 presents the experimental results using box plots. Fig. 5(a) and (b) show the three-dimensional scatter plots in terms of test accuracy, rule number, and gene number for data sets lung cancer and SRBCT, respectively.

From Table 2, we can observe that iGEC performs better than the Vinterbo's classifier using 200 candidate genes (V200) in the five measures: TrCR (97.1% versus 81.5%), TeCR (87.9% versus 81.2%), N_r (3.9 versus 4.9), N_f (5.0 versus 7.2), and N_r/C (1.1 versus 1.4). Note that V200 is better than V15 but using more candidate genes and computation time. Moreover, the classifiers V200 compare favorably to those of a logistic regression model which is one of the frequently used classification method applied in the biomedical domain (Vinterbo et al., 2005).

Fig. 6 shows an example of iGEC using the data set leukemia1 where 90% samples are for training and the rest for test. The classifier has four fuzzy rules using three genes L05148, U46499, and U05259, where TrCR = 100% and TeCR = 100%. The fuzzy rules are linguistically interpretable as follows:

R_1 If L05148 is not up-regulated and U05259 is not down-regulated, then class "ALL B-Cell" with $CF = 0.243$;
R_2 If L05148 is ALL and U46499 is neutral or up-regulated, then class "ALL B-Cell" with $CF = 0.682$;
R_3 If L05148 is not down-regulated, U46499 is ALL and U05259 is ALL, then class "ALL T-Cell" with $CF = 0.710$;
R_4 If L05148 is ALL, U46499 is ALL and U05259 is ALL, then class "AML" with $CF = 0.722$.

Where the membership functions of genes U46499 and U05259 in R_1 and R_2, respectively, are "*don't care*" which can reduce the rule length. From the compact rule base, it is easy to interpret the classification model from gene expression data. The fuzzy rules can be examined by biomedical researchers. Due to the natural clustering property of gene expression data, each of the classes "ALL T-Cell" and "AML" has one fuzzy rule corresponding to one fuzzy region while the class "ALL B-Cell" has two fuzzy regions overlapped. Furthermore, we can know the distribution of samples of each class from the corresponding membership function in the feature space. The fuzzy rule base can determine the class of unknown samples using Eq. (2).

To further realize whether these three genes L05148, U46499, and U05259 make sense as a group and their biological relationship, we process the average link-

Table 2
The statistical results of iGEC and the Vinterbo's classifier on training accuracy (TrCR), test accuracy (TeCR), number of rules (N_r), number of genes (N_f), and rule number per class (N_r/C)

Data set	Method	TrCR (%)	TeCR (%)	N_r	N_f	N_r/C
Brain tumor1	iGEC	92.4 ± 0.1	88.7 ± 2.5	5.0 ± 0.1	5.9 ± 0.2	1.00
	V200	80.85	81.25	6.50	8.60	1.30
	V15	78.66	85.00	6.00	9.20	1.20
Brain tumor2	iGEC	97.0 ± 0.2	72.4 ± 4.4	4.4 ± 0.1	5.5 ± 0.3	1.11
	V200	60.00	60.00	4.00	8.30	1.00
	V15	66.60	63.33	5.10	6.70	1.27
DLBCL	iGEC	98.5 ± 0.7	91.2 ± 1.8	2.5 ± 0.3	3.7 ± 0.4	1.28
	V200	85.91	85.00	2.60	3.80	1.30
	V15	84.65	78.33	7.00	6.90	3.50
Leukemia1	iGEC	99.7 ± 0.2	94.0 ± 2.5	3.5 ± 0.1	4.1 ± 0.3	1.18
	V200	90.15	92.00	5.30	7.30	1.76
	V15	87.61	84.00	4.90	8.10	1.63
Leukemia2	iGEC	98.7 ± 0.3	85.3 ± 2.7	3.3 ± 0.1	4.3 ± 0.3	1.12
	V200	81.97	76.67	4.30	5.50	1.43
	V15	74.70	71.67	3.50	4.10	1.16
Lung cancer	iGEC	92.7 ± 0.2	88.0 ± 2.7	5.5 ± 0.2	6.9 ± 0.4	1.10
	V200	85.35	84.44	7.80	14.50	1.56
	V15	81.57	82.78	8.30	8.90	1.66
Prostate tumor	iGEC	97.9 ± 0.5	90.9 ± 4.0	2.4 ± 0.2	4.1 ± 0.4	1.21
	V200	81.5	82.00	3.00	3.30	1.50
	V15	84.46	84.00	2.90	5.10	1.45
SRBCT	iGEC	99.8 ± 0.5	92.3 ± 9.9	4.3 ± 0.2	4.8 ± 0.3	1.08
	V200	86.36	88.33	5.80	6.20	1.45
	V15	78.44	71.67	5.10	10.20	1.27
Mean	iGEC	97.1	87.9	3.9	5.0	1.1
	V200	81.5	81.2	4.9	7.2	1.4
	V15	79.6	77.6	5.4	7.4	1.6

age (average distance, UPGMA) clustering based on Euclidean distances squared by EPCLUST (Parkinson et al., 2003). Fig. 7 shows the clustering result. From Fig. 7, we can observe that most of the samples belonging to same class are grouped together. From thousands of genes, the proposed method can identify few but relevant genes to make accurate classification. Furthermore, the biological finding is interpretable from the obtained compact fuzzy rule base. Therefore, iGEC is beneficial to microarray data analysis and development of inexpensive diagnostic tests.

Besides the leukemia1 classifier using the gene set {L05148, U46499, U05259} shown in Fig. 6, there are other sets of three genes which can establish the classifiers with both 100% training and test accuracies as follows: {L05148, M63138, U05259}, {M11722, L05148, U46499}, {M31523, U16954, U46499}, and {U16954, M27891, U05259}. This scenario results from that the microarray data have a large number of genes but a very small number of samples. iGEC can provide important knowledge to biological scientists. Table 3 gives descriptions of the selected genes from the data set leukemia1 of 72 samples. For each gene, we counted the number of articles that were retrieved by a PubMed query containing the gene name and the string "leukemia". By combining more gene sets of solutions, most of genes highly related to the leukemia disease can be obtained.

Due to different merits of fuzzy partitions such as grid partition, tree partition, and scatter partition, they cannot be directly compared using some specific measurements (Ho et al., 2004a). However, iGEC has 1.1 fuzzy regions for describing the sample distribution of each class averagely. Besides the above-mentioned advantages of easy interpretation and economical experiments, the proposed fuzzy rule-base method using a scatter partition of feature space can enclose all possible occurrences of samples in the same class with one or few hyperbox-type

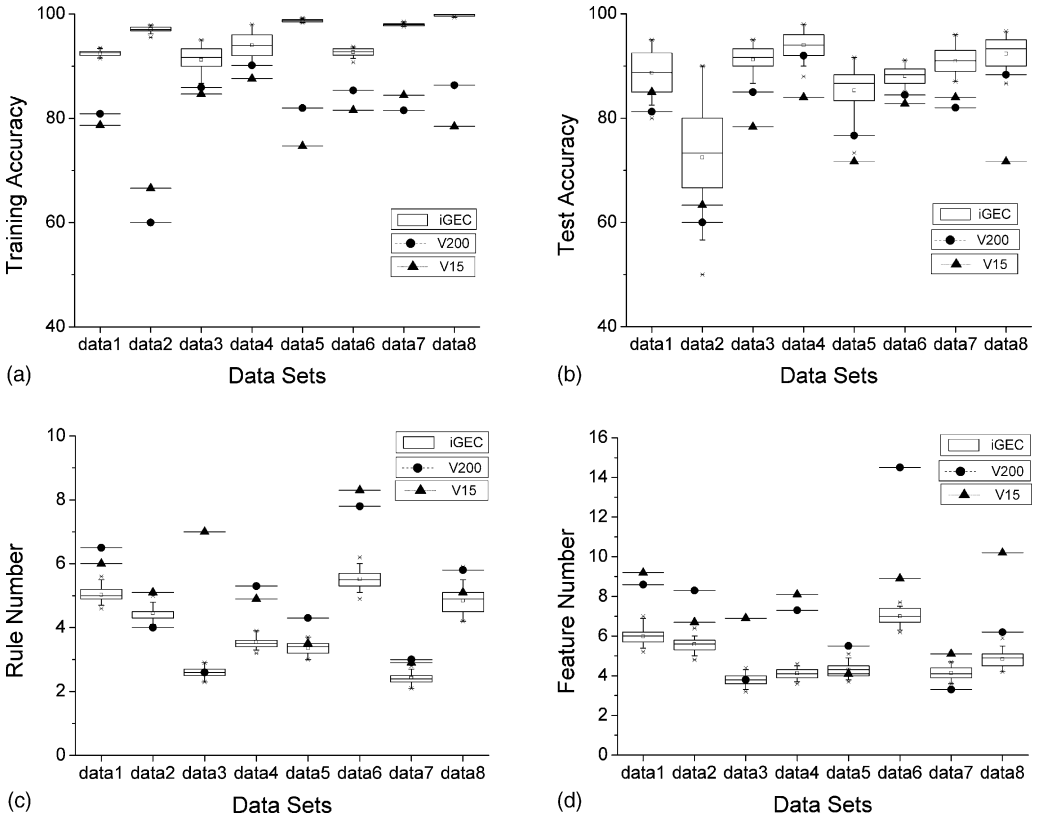

Fig. 4. The box plots of the statistical results: (a) training accuracy; (b) test accuracy; (c) number of rules and (d) number of used genes.

fuzzy regions. In other words, the fuzzy regions of scatter partition can represent one class more independently than those of grid partition. Therefore, iGEC can reject the unknown sample if it belongs to no fuzzy region that no fuzzy rule is fired.

To further evaluate accuracy of the proposed method, we compared iGEC with some non-rule-based classifiers without using gene selection methods in Statnikov et al. (2005). Table 4 shows the test accuracy comparisons using 10-CV on the eight data sets between iGEC and the following methods: multi-category support vector machine (SVM), *k*-nearest neighbors (*k*-NN), backpropagation neural networks (NN), and probabilistic neural networks (PNN) which are the most common methods for gene expression data analysis. The results are obtained from Statnikov et al. (2005).

Table 4 indicates that the multi-category SVM with 93.63% average test accuracy on the eight data sets is the

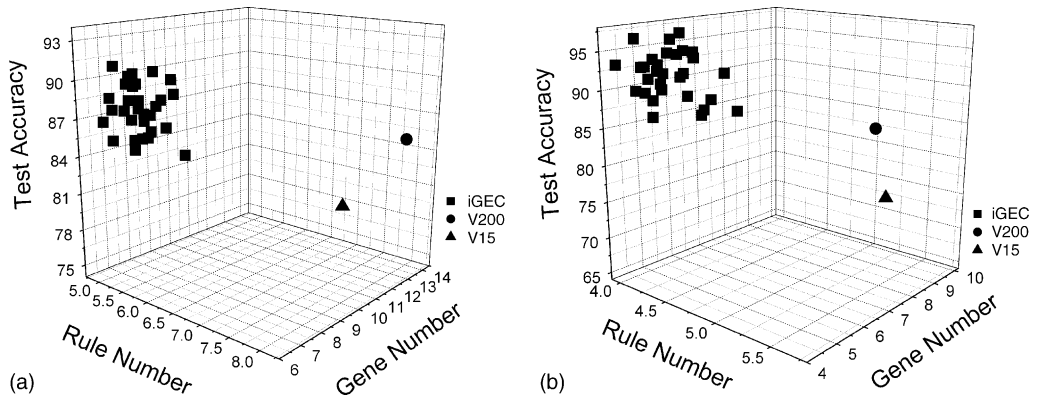

Fig. 5. The 3D scatter plots: (a) lung cancer and (b) SRBCT.

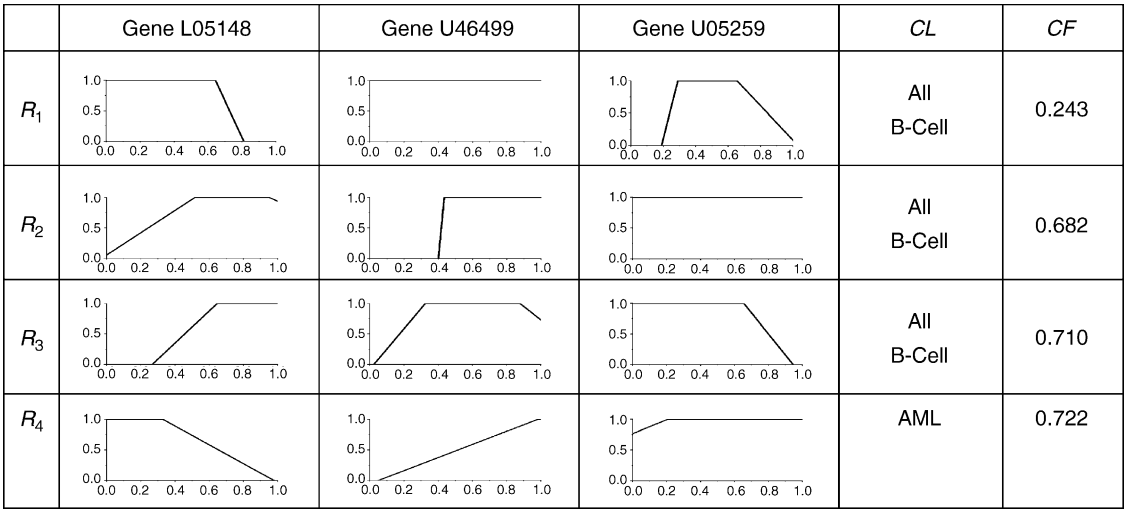

Fig. 6. Fuzzy rules of the data set leukemia1 using 90% samples for training and the rest for test. The training and test accuracies are both 100%.

Table 3
Selected genes for the leukemia1 data set example

Gene	Description	No. of references
M11722	Human terminal transferase mRNA	154
L05148	Human protein tyrosine kinase related mRNA sequence	26
M63138	Human cathepsin D	24
M31523	Human transcription factor (E2A) mRNA	17
U05259	Human MB-1 gene, complete cds	12
U46499	Homo sapiens microsomal glutathione transferase (MGST1) gene, 3′ sequence	10
M27891	Human cystatin C gene	5
U16954	Human (AF1q) mRNA	3

For each gene we counted the number of articles that were retrieved by a PubMed query consisting of the gene name and the string "leukemia".

most accurate classifier for diseases classification. However, it is not practical to use as many as 7965.6 genes on average to classify diseases samples for economical biomedical test in real applications. The proposed fuzzy classifier iGEC with 87.9% using 5.0 genes on average is superior to *k*-NN (84.49%), NN (82.54%), and PNN (79.49%) in terms of accuracy and number of genes. Because the sample sizes of microarray data are extremely small, it results in the high training accuracy (97.1%) and relatively low test accuracy (87.9%). When the number of samples is increased, the test accuracy can be further advanced (Ho et al., 2004a). From the viewpoint of analysis and practical applications, iGEC can serve as one of efficient tools for analysis of gene expression profiles.

Table 4
The test accuracies and numbers of used genes for iGEC and non-rule-based classifiers using 10-CV

Data set	No. of genes in non-rule-based classifiers	Accuracy (%)					No. of genes in iGEC
		SVM	*k*-NN	NN	PNN	iGEC	
Brain tumor1	5920	91.67	87.94	84.72	79.61	88.71	6
Brain tumor2	10367	77.00	68.67	60.33	62.83	72.45	6
DLBCL	5469	97.50	86.96	89.64	80.89	91.22	4
Leukemia1	5327	97.50	83.57	76.61	85.00	94.00	4
Leukemia2	11225	97.32	87.14	91.03	83.21	85.33	4
Lung cancer	12600	96.05	89.64	87.80	85.66	88.09	7
Prostate tumor	10509	92.00	85.09	79.18	79.18	90.97	4
SRBCT	2308	100.00	86.90	91.03	79.50	92.33	5
Mean	7965.6	93.63	84.49	82.54	79.49	87.89	5.0

The results of the non-rule-based classifiers without using gene selection methods are obtained from Statnikov et al. (2005).

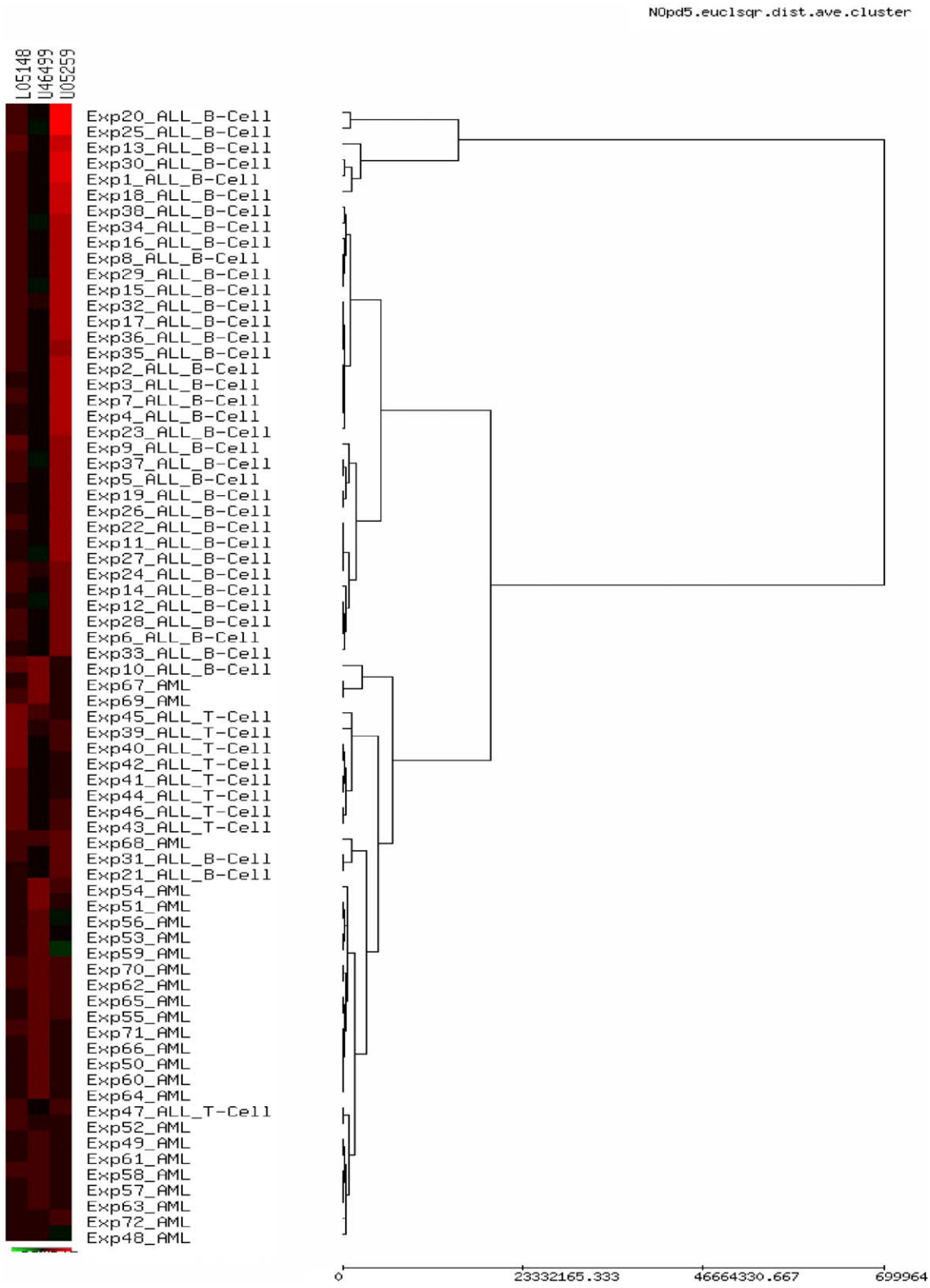

Fig. 7. The clustering result of 72 samples in data set leukemia1 using the three selected genes by the clustering algorithm EPCLUST (Parkinson et al., 2003).

4. Conclusions

Microarray data analysis and gene expression classification are important research topics in bioinformatics such that how to design an accurate, compact, and linguistically interpretable classifier is the major concern in this study. We proposed an interpretable gene expression classifier, named iGEC, for microarray data analysis. The design of iGEC includes almost all aspects related to the design of compact fuzzy rule-based classification systems: gene selection, rule selection, membership function tuning, consequent class determination, and certainty grade tuning. Consequently, an efficient optimization algorithm IGA is used to solve the resultant optimization problem with a large number of parameters.

The superiority of the proposed iGEC was evaluated by computer simulation on eight data sets of gene expression. The experimental results reveal that the proposed method can obtain interpretable classifiers with an accurate and compact fuzzy rule base, compared with the existing fuzzy classifier. iGEC is an efficient tool for analysis of gene expression profiles.

Acknowledgements

The authors would like to thank S.A. Vinterbo and the co-authors for providing the program of their method.

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.

Creighton, C., Hanash, S., 2003. Mining gene expression databases for association rules. Bioinformatics 19 (1), 79–86.

Deb, K., Reddy, A.R., 2003. Reliable classification of two-class cancer data using evolutionary algorithms. BioSystems 72, 111–129.

Ding, C., 2003. Unsupervised feature selection via two-way ordering in gene expression analysis. Bioinformatics 19 (10), 1259–1266.

Goldberg, D.E., 1989. Genetic Algorithms in Search. Optimization and Machine Learning. Addison-Wesley, Reading, MA.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., 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., Chen, H.-M., Ho, S.-J., Chen, T.-K., 2004a. Design of accurate classifiers with a compact fuzzy-rule base using an evolutionary scatter partition of feature space. IEEE Trans. Systems Man Cybern. Part B 34 (2), 1031–1044.

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

Hvidsten, T.R., Lgreid, A., Komorowski, J., 2003. Learning rulebased models of biological process from gene expression time profiles using gene ontology. Bioinformatics 19 (9), 1116–1123.

Kauffman, S., Peterson, C., Samuelsson, B., Troein, C., 2003. Random Boolean network models and the yeast transcriptional network. PNAS 100 (25), 14796–14799.

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.

Li, J., Liu, H., Downing, J.R., Yeoh, A.E.-J., Wong, L., 2003. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (all) patients. Bioinformatics 19 (1), 71–78.

Li, L., Clarice, 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.

Liu, B., Cui, Q., Jiang, T., Ma, S., 2004. A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC Bioinform. 5, 136–147.

Liu, X., Krishnan, A., Adrian Mondry, A., 2005a. An entropy-based gene selection method for cancer classification using microarray data. BMC Bioinform. 6 (1), 76–89.

Liu, J., Cutler, G., Li, W., Pan, Z., Peng, S., Hoey, T., Chen, L., Ling, X., 2005b. Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics 21 (11), 2691–2697.

Merz, P., 2003. Analysis of gene expression profiles: an application of memetic algorithms to the minimum sum-of-squares clustering problem. BioSystems 72, 99–109.

Michalewicz, Z., Dasgupta, D., Le Riche, R.G., Schoenauer, M., 1996. Evolutionary algorithms for constrained engineering problems. Comput. Ind. Eng. 30 (4), 851–870.

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.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Garcia Lara, G., Holloway, E., Kapushesky, M., Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-Serra, P., Sharma, A., Sansone, S., Brazma, A., 2003. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 31 (1), 68–71.

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.

Ressom, H., Reynolds, R., Varghese, R.S., 2003. Increasing the efficiency of fuzzy logic-based gene expression data analysis. Physiol. Genomics 13, 107–117.

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.

Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D., Levy, S., 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21, 631–643.

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

Wahde, M., Hertz, J., 2000. Coarse-grained reverse engineering of genetic regulatory networks. BioSystems 55, 129–136.

Woolf, P.J., Wang, Y., 2000. A fuzzy logic approach to analyzing gene expression data. Physiol. Genomics 3, 9–15.

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.

Zhou, X., Mao, K.Z., 2005. LS bound based gene selection for DNA microarray data. Bioinformatics 21 (8), 1559–1564.