Artificial Intelligence in Medicine (2008) **42**, 81–93

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer

Zne-Jung Lee *

Department of Information Management, Huafan University, No. 1, Huafan Road, Shihding Township, Taipei County 22301, Taiwan, ROC

Received 29 January 2007; received in revised form 27 September 2007; accepted 27 September 2007

KEYWORDS
Ovarian cancer;
Microarray data;
Gene selection;
Support vector machine;
Genetic algorithm;
Particle swarm optimization

Summary

Objective: The type of data in microarray provides unprecedented amount of data. A typical microarray data of ovarian cancer consists of the expressions of tens of thousands of genes on a genomic scale, and there is no systematic procedure to analyze this information instantaneously. To avoid higher computational complexity, it needs to select the most likely differentially expressed gene markers to explain the effects of ovarian cancer. Traditionally, gene markers are selected by ranking genes according to statistics or machine learning algorithms. In this paper, an integrated algorithm is derived for gene selection and classification in microarray data of ovarian cancer.
Methods: First, regression analysis is applied to find target genes. Genetic algorithm (GA), particle swarm optimization (PSO), support vector machine (SVM), and analysis of variance (ANOVA) are hybridized to select gene markers from target genes. Finally, the improved fuzzy model is applied to classify cancer tissues.
Results: The microarray data of ovarian cancer, obtained from China Medical University Hospital, is used to test the performance of the proposed algorithm. In simulation, 200 target genes are obtained after regression analysis and six gene markers are selected from the hybrid process of GA, PCO, SVM and ANOVA. Additionally, these gene markers are used to classify cancer tissues.
Conclusions: The proposed algorithm can be used to analyze gene expressions and has superior performance in microarray data of ovarian cancer, and it can be performed on other studies for cancer diagnosis.

* Tel.: +886 2 26632102#4356; fax: +886 2 26632102#4353.
E-mail address: johnlee@hfu.edu.tw.

0933-3657/$ – see front matter
doi:10.1016/j.artmed.2007.09.004

1. Introduction

Ovarian cancer is classified as the epithelial cancer caused by genetic alterations that disrupts regulation of proliferation, apoptosis, senescence and DNA repair [1]. It is the primary cause of gynecological cancer death for women in the United States [2—4]. Thus, early detection is paramount for ovarian cancer, because 70% of women with the epithelial ovarian cancer are not diagnosed until the disease is spread to upper abdomen (stage III) [4]. Recently, microarray technology is a newly emerging technique of gene chip and becomes a common tool to globally analyze tens of thousands of genes on a genomic scale. In general, the layout of microarray data is an ordered array of microscopic element on a planar substrate which contains the genotype of many relevant or irrelevant genes to cancer development. A microarray data is a device that can be employed in comparison of gene expression level, patient genotype and development of new medicine, etc. It can provide useful biological, diagnostic and prognostic information for researchers [5]. This means that this technology enables biologists to investigate and compare the expression level of genes under variant conditions [6,7]. In this study, microarray technology is applied for the real data obtained from China Medical University Hospital.

Typically, the microarray data of ovarian cancer contains genes with tens of thousands of dimension and only contain few samples [3]. It is a non-trivial task to tackle the genes with such a high dimension. Besides, it is difficult to categorize ovarian cancer on the basis of their genes expressions when the size of samples is small [7—10]. Recently, there are many methods such as feature selection, correlation methods, nonparametric scoring approach, and Bayesian variable selection approach that are proposed to select informative genes from microarray data [7,11—23]. But there is no consideration about the variable gene selection and block effect in microarray data. A number of machine learning methods and genomic expression for ovarian cancers have been successfully applied to the analysis of gene expression in microarray data [24—29]. A key emphasis in above literature is to identify gene products that could act as specific markers of ovarian cancer, but the time complexity is rather high due to genomic expression with high dimension. Additionally, there is no systematic approach for achieving a better insight into global gene expression analysis and no convincing new marker has been identified [3]. In this paper, an integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer is proposed. In the proposed algorithm, regression analysis is first performed to find target genes in the proposed algorithm. A hybrid process based on genetic algorithm, particle swarm optimization, support vector machine and analysis of variance are conducted to find gene markers from target genes. Furthermore, the improved fuzzy model is applied to verify cancer tissues by the selected gene markers.

The rest of this paper is organized as follows: Section 2 will review these important methods needed for the proposed algorithm. In Section 3, the proposed algorithm is discussed in detail. The used microarray data is also introduced in this section. The simulation results will presented in Section 4. Section 5 contains conclusions and direction of future work.

2. Methods

Here a brief description of important methods in the proposed algorithm such as genetic algorithm, particle swarm optimization and support vector machine is given.

2.1. Genetic algorithm

Genetic algorithm has been touted as a class of general-purpose search strategies for optimization problems [30]. GA can handle any kind of objective functions and any kind of constraints without much mathematical requirements about the optimization problems [31—33]. In GA variables of a problem are represented as genes in a chromosome, and the chromosomes are evaluated according to their fitness using some measures of profit or utility that we want to optimize. The recombination typically involves two genetic operators: crossover and mutation. The genetic operators alter the composition of genes to create new chromosomes called offspring. In the crossover operator, it generates offspring that inherits genes from both parents with a crossover probability P_c. In the mutation operator, it inverts randomly chosen genes in a chromosome with a mutation probability P_m. The selection operator is modeled from natural selection that fitter chromosomes survive and weaker ones die. The fitter chromosome has higher probabilities of being selected in the next generation. After several generations, GA can converge to the best solution.

2.2. Particle swarm optimization

Kennedy and Elberhart proposed PSO that is based on metaphor of social interaction and communica-

tion such as bird flocking and fish schooling [34]. In PSO, each member is called a particle. A particle represents a potential solution, and each particle has a position and velocity vector. A population of particles is randomly generated initially, and then performs the search for the optima iteratively. The position vector X_i^n and velocity vector V_i^n in the n-dimension of the ith particle can be represented as $P_i = (P_i^1, P_i^2, \ldots, P_i^n)$ and $V_i = (V_i^1, V_i^2, \ldots, V_i^n)$ respectively. The PSO algorithm is described as follows [35–38]:

$$V_i^n = \beta \times V_i^n + c1 \times \text{rand}() \times (\text{pbest}_i^n - P_i^n) + c2 \times \text{rand}() \times (\text{gbest}^n - P_i^n) \quad (1)$$

$$P_i^n = P_i^n + V_i^n \quad (2)$$

where β denotes the inertia weight, $c1$ and $c2$ are the acceleration constants, rand() the uniformly distributed random number between 0 and 1, $\text{pbest}_i = (\text{pbest}_i^1, \text{pbest}_i^2, \ldots, \text{pbest}_i^n)$ the best previous position yielding the best object function for the ith particle, and gbest = (gbest^1, gbest^2, …, gbest^n) is the best position discovered by the whole population. The inertia weight, β is a user-defined parameter that controls, with $c1$ and $c2$, the previous values of particle velocities on it current one. The particle's new velocity is calculated by Eq. (1) according to its previous velocity and to the distances of its current position from its own best historical position and the collaborative effect of particles. Then, the particle updates the new position according to Eq. (2).

2.3. Support vector machine

SVM is a new classification technique originally proposed by Vapnik, and successively applied to many fields [39]. The SVM technique is briefly described as follows. Let (x_i, y_i), $1 \le i \le N$, be a set of training examples, where N is the number of training data. For each sample data, it must conform to the following:

$$x_i \in R^n \quad \text{and} \quad y_i \in \{-1, +1\} \quad (3)$$

where n is the number of dimensions of input data. The objective of SVM is to find an optimal separating hyper-plane with the maximum margin (w) and a real value b for classification of data. In practice, the data may not be linearly separable, and some classification errors are allowed on the learning sets. This problem is done by its soft margin version, i.e. by the problem:

$$\underset{w,b}{\text{Min}} \ \frac{1}{2} w^{\mathrm{T}} w + C \sum_{i=1}^{N} \xi_i \quad (4)$$

subject to

$$y_i(\langle w \cdot x_i \rangle + b) \ge 1 - \xi_i, \qquad \xi_i \ge 0 \quad \text{and} \quad i = 1, \ldots, N \quad (5)$$

where the slack variables $\xi_i \ge 0$, $i = 1, \ldots, N$. The above optimization model can also be solved using Lagrange method as follows:

$$\underset{\alpha}{\text{Max}} \ L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i \cdot x_j \rangle \quad (6)$$

subject to

$$0 \le \alpha_i \le C, \qquad i = 1, \ldots, N \quad \text{and} \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \quad (7)$$

where C is the upper bound on α_i.

Some classification problems do not have a satisfactory solution but have a nonlinear one. The nonlinear SVM can map the input data into a high-dimension feature space via a mapping function $\phi(x)$. By virtue of constructing the feature space, it can substitute $\phi(x)$ into Eq. (6) and has the following:

$$\underset{\alpha}{\text{Max}} \ L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle \quad (8)$$

Given a kernel function $k(x, y) = \langle \phi(x), \phi(y) \rangle$, Eq. (8) is showed as follows:

$$\underset{\alpha}{\text{Max}} \ L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \quad (9)$$

For Gaussian radial basis function kernel (GRBF), the kernel function is defined as

$$k(x_i, x_j) = \exp\left(- \frac{||x_i - x_j||^2}{2\sigma} \right)$$

3. The used microarray data and the proposed algorithm

The ovarian tissues, vaginal tissues, cervical tissues and myometrium from patients were collected at China Medical University Hospital. Tissues applied in this study include five ovarian tumors (OVT), five ovarian cancers at stage I (OVCAI) and five ovarian cancers at stage III (OVCAIII). The clinical-pathologic characteristics of OVCAI and OVCAIII are shown in Table 1 [3]. All microarrary procedures were also performed in a dust/climate control laboratory at China Medical University. A sequence-verified human cDNA library containing 9600 human cDNA clones was a kind gift from the National Health Research Institute of Taiwan. The clones were originally

Table 1 The clinical-pathologic characteristics for ovarian cancers

Age	Pathology	Stage
44	Mucinous cytadenoma, borderline malignancy	IA
39	Serous cytadenoma, borderline malignancy	IA
46	Mucinous cytadenoma, borderline malignancy	IA
26	Mucinous cytadenoma, borderline malignancy	IIIB
36	Mucinous cytadenoma, borderline malignancy	IA
51	Serous cystadenocarcinoma	IIIC
58	Clear cell carcinoma	IC
62	Serous cystadenocarcinoma	IIIC
53	Clear cell carcinoma; endometrioid adenocarcinoma	IIIC
61	Serous cystadenocarcinoma	IIIC

obtained from the minimum information about a microarray experiment (MIAME) consortium libraries through its distributor (Research Genetics, Huntsville, AL) [42].

In this paper, an integrated algorithm for gene selection and classification in microarray data of ovarian cancer is proposed. The flowchart of the proposed algorithm is shown in Fig. 1. In regression analysis, it is used to reduce the dimension (9600 genes) and then to find target genes. The ratio of Ch1/Ch2 based on the background corrected mean fluorescence intensity of the (red) pixels is used to perform analysis. It can be expressed as the following form:

$$\frac{\mathrm{Ch1}}{\mathrm{Ch2}} = \frac{\mathrm{Ch1I} - \mathrm{Ch1B}}{\mathrm{Ch2I} - \mathrm{Ch2B}} = \frac{X}{Y} \tag{10}$$

where Ch1I and Ch2I are the uncorrected mean pixel intensities in the microarray data. Ch1B and Ch2B are the median intensities of the background pixels [43]. Regression analysis calculates for all X and Y samples. The residual of every gene represents the errors of groups, and the errors of groups consist of the sum of all vertical distances from the regression line [44]. The residuals of 9600 genes after regression analysis are shown in Fig. 2. After regression analysis, 100 genes with maximum residuals and 100 genes with minimum residuals are selected as target genes. The genes names of these 200 target genes are listed in Table 2. Furthermore, a hybrid process of GA, PSO, SVM and ANOVA are applied to select gene markers. To implement this hybrid process, the RBF kernel function is used for SVM. The RBF kernel function is suitable for high-dimensional data and only needs to define two parameters (C and σ) [45]. In order to find the best values of C and σ, both parameters are encoded in the chromosome of GA and particles of PSO. For initialization, the initial population of GA and particles of PSO are generated randomly. For GA and PSO, chromosomes and particles are encoded as gene masks corresponding to all human cDNA clones, C and σ. For GA, the binary value of each gene mask with 1 represents that this gene is selected for verifying OVT and OVCA, and 0 indicates that this gene is not selected. For PSO, a sigmoid function is defined as follows:

$$S(P_i^n) = \frac{1}{1 + \mathrm{e}^{-P_i^n}} \quad \text{if } (\mathrm{rand}() < S(P_i^n)) \quad \text{then } P_i^n = 1; \quad \text{else } P_i^n = 0 \tag{11}$$

where $P_i^n = 1$ represents that the gene i is selected otherwise this gene i is not selected. For GA, the values of C and σ are encoded into binary string and the mapping from a binary string to a real number (r) is calculated as follows:

$$r = \min_r + \mathrm{binrep} \times \frac{\max_r - \min_r}{2^l - 1} \tag{12}$$

where $\min_r$ is the minimum value of C and σ, $\max_r$ the maximum value of C and σ, and binrep represents the decimal value of l length. For PSO, the values of C and σ are encoded into as real variables between $\min_r$ and $\max_r$. For GA, the two-point crossover will act on parents to generate offspring. Mutation is a random alternation of a bit in the chromosome which assists in keeping diversity in the population. The traditional roulette selection with elitism is performed as selection operator, and it ensures that the best chromosome is selected into the new generation. For PSO, new particles are created by Eqs. (1) and (2). The inertia weight, β is given by

$$\beta = (\beta_1 - \beta_2) \times \frac{(\mathrm{MAXGEN} - \mathrm{gen})}{\mathrm{MAXGEN}} + \beta_2$$

where β_1 and β_2 are the initial and final value of weight, respectively, gen the current generation number and MAXGEN is the maximum number of generation. In the hybrid process, SVM is applied to classify cancer tissues. There are 12 samples used as training data and 3 samples are randomly

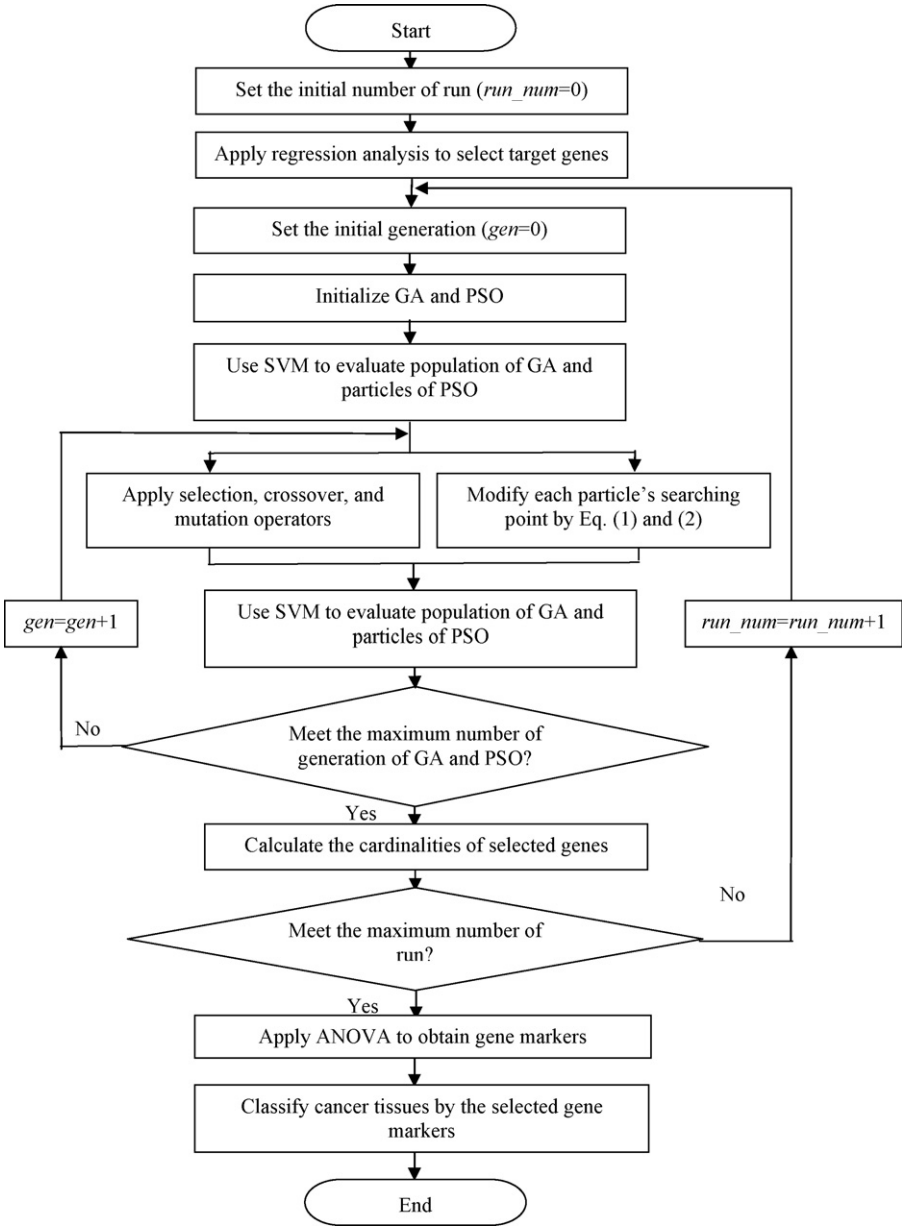

Figure 1 The flowchart of the proposed algorithm.

selected to test the accuracy of classification. For GA and PSO, the fitness is calculated as follows:

$$\text{fitness} = W_a \times \text{SVM}_a + W_b \times \left(N - \sum_{i=1}^{\lambda} G_i \right) \qquad (13)$$

where W_a is the weight for the classification accuracy of SVM, SVM_a the classification accuracy of SVM, W_b the weight for the selected gene, λ the number of target genes, $G_i = 1$ represents that the gene i is selected, and $G_i = 0$ indicates that this gene i is not selected. In the hybrid process, the cardinalities of selected genes are calculated. It is noted that GA and PSO are simultaneously processed with SVM [48,49]. The cardinalities of selected target genes from both GA and PSO with SVM are instantaneously accumulated. This process is repeated until the maximum number of generation has satisfied. Thereafter, ANOVA is performed to select gene markers. ANOVA can measure the difference between the means of two groups and can be used to discover the difference between OVT and OVCA [46]. In the proposed algorithm, ANOVA is applied to select gene markers from the 50 target genes with the highest cardinalities. The obtained gene markers are furthermore verified by the improved fuzzy model. In the improved fuzzy model, there is a set of observation $\{(\vec{x}(1), y_1), (\vec{x}(2), y_2), \ldots, (\vec{x}(N), y_N)\}$ with $\vec{x}(i) \in R^n$ and $y_i \in R$. Where $\vec{x}(i) = [x_1(i),$

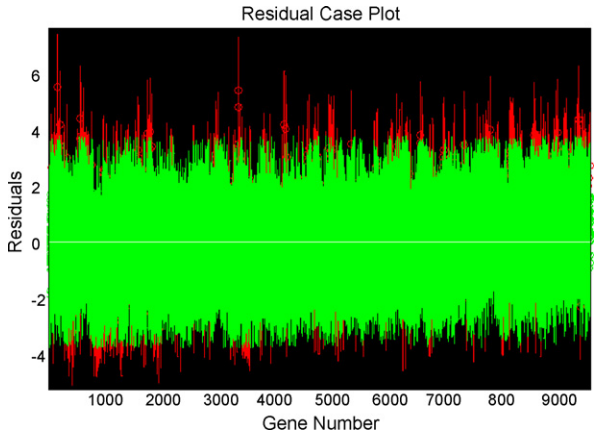

Figure 2 The residuals of 9600 genes after regression analysis.

$x_2(i), \ldots, x_n(i)]$ is the ith input vector and $x_j(i)$ the input, y_i the desired output, and N is the number of training data. The improved fuzzy model consists of IF-THEN rules that can be expressed as the following form:

$$R^i: \text{If } x_1(i) \text{ is } A_1^i(\Omega_1^i, x_1(i)) \quad \text{and}$$
$$x_2(i) \text{ is } A_2^i(\Omega_2^i, x_2(i)), \ldots, x_n(i) \text{ is } A_n^i(\Omega_n^i, x_n(i))$$
$$\text{then } y^i = f_i(x_1(i), x_2(i), \ldots, x_n(i); \vec{a}^i)$$
$$= 1 + a_1^i x_1(i) + \cdots + a_n^i x_n(i) \tag{14}$$

where $R^i (i = 1, 2, \ldots, C)$ is the ith fuzzy rule, C the numbers of fuzzy rules, $A_j^i(\Omega_j^i, x_j(i))$ the fuzzy set of the ith rule with the parameter set Ω_j^i, y^i the output of the fuzzy rule R^i, and $\vec{a}^i = (1, a_1^i, \ldots, a_n^i)$ is the

Table 2 The gene names of 200 target genes

100 genes with minimum residuals	
1	'nuclear receptor subfamily 4, group A, member 1'
2	'early growth response 1'
3	'FBJ murine osteosarcoma viral oncogene homolog B'
4	2764
5	'jun B proto-oncogene'
6	'isocitrate dehydrogenase 3 (NAD+) alpha'
7	'selenoprotein P, plasma, 1'
8	'v-jun avian sarcoma virus 17 oncogene homolog'
9	'nuclear receptor subfamily 4, group A, member 3'
10	'jun B proto-oncogene'
11	'regulator of G-protein signalling 2, 24 kD'
12	'complement component 1, s subcomponent'
13	'jun D proto-oncogene'
14	'serum/glucocorticoid regulated kinase'
15	'jun B proto-oncogene'
16	'poly(A)-binding protein, cytoplasmic 1-like'
17	'matrix metalloproteinase 2 (gelatinase A, 72 kDa gelatinase, 72 kDa type IV collagenase)'
18	'Human insulin-like growth factor binding protein 5 (IGFBP5) mRNA'
19	'transient receptor potential channel 3'
20	'KIAA1053 protein'
21	'vimentin'
22	'Kruppel-like factor 4 (gut)'
23	'nuclear receptor subfamily 0, group B, member 1'
24	'ras homolog gene family, member B'
25	'jun D proto-oncogene'
26	'ras homolog gene family, member B'
27	'SPARC-like 1 (mast9, hevin)'
28	'TED protein'
29	'KIAA1064 protein'
30	'electron-transfer-flavoprotein, alpha polypeptide (glutaric aciduria II)'
31	'collagen, type I, alpha 2'
32	'ribosomal protein L6'
33	'v-jun avian sarcoma virus 17 oncogene homolog'
34	'distal-less homeobox 4'
35	'*Homo sapiens* mRNA; cDNA DKFZp434A115 (from clone DKFZp434A115)'
36	'myosin, heavy polypeptide 9, non-muscle'
37	'adrenomedullin'
38	'HMT1 (hnRNP methyltransferase, *S. cerevisiae*)-like 2'
39	'insulin induced protein 2'

Table 2 (*Continued*)

40	'major histocompatibility complex, class I, A'
41	'*Homo sapiens* mRNA; cDNA DKFZp564E026 (from clone DKFZp564E026)'
42	'RNA binding motif protein 4'
43	'potassium large conductance calcium-activated channel, subfamily M, beta member 3-like'
44	'transforming growth factor beta-stimulated protein TSC-22'
45	'transcription factor 6-like 1 (mitochondrial transcription factor 1-like)'
46	'ribosomal protein L32'
47	'complement component 1, r subcomponent'
48	'inositol polyphosphate-4-phosphatase, type II, 105 kDa'
49	'myosin regulatory light chain 2, smooth muscle isoform'
50	'inhibitor of DNA binding 2, dominant negative helix-loop-helix protein'
51	'heat shock 70 kDa protein 8'
52	'suppression of tumorigenicity 13 (colon carcinoma) (Hsp70-interacting protein)'
53	'ESTs'
54	'inhibitor of DNA binding 2, dominant negative helix-loop-helix protein'
55	'microtubule-associated protein, RP/EB family, member 3'
56	'integral membrane protein 2B'
57	'matrix metalloproteinase 16 (membrane-inserted)'
58	'minichromosome maintenance deficient (*S. cerevisiae*) 7'
59	'RNA helicase-related protein'
60	'a disintegrin-like and metalloprotease (reprolysin type) with thrombospondin type 1 motif, 4'
61	'transcription elongation factor A (SII)-like 1'
62	'KIAA0173 gene product'
63	'immunoglobulin lambda-like polypeptide 1'
64	'ESTs'
65	8256
66	'leptin receptor gene-related protein'
67	'collagen, type III, alpha 1 (Ehlers—Danlos syndrome type IV, autosomal dominant)'
68	'general transcription factor IIH, polypeptide 2 (44 kDa subunit)'
69	'nuclear receptor subfamily 4, group A, member 2'
70	'v-akt murine thymoma viral oncogene homolog 2'
71	'ribosomal protein S4, X-linked'
72	'serum/glucocorticoid regulated kinase'
73	'GL002 protein'
74	'*Homo sapiens* clone 23892 mRNA sequence'
75	'golgi transport complex 1 (90 kDa subunit)'
76	'EST'
77	'guanine nucleotide binding protein 11'
78	'*Homo sapiens*, clone MGC:5618, mRNA, complete cds'
79	'lymphocyte activation-associated protein'
80	'T-box, brain, 1'
81	'DKFZP566D213 protein'
82	'*Homo sapiens* cDNA FLJ10205 fis, clone HEMBA1004954'
83	'olfactory receptor, family 2, subfamily A, member 7'
84	'nuclear receptor subfamily 2, group F, member 2'
85	'dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2'
86	'complement component 7'
87	'AXL receptor tyrosine kinase'
88	'four and a half LIM domains 2'
89	'phosphoglycerate dehydrogenase'
90	'aldehyde dehydrogenase 1 family, member A1'
91	'RNA binding protein; AT-rich element binding factor'
92	'hypothetical protein FLJ20160'
93	'peptidyl arginine deiminase, type II'
94	'motilin'
95	806
96	'BCL2-like 1'
97	'cullin 5'
98	'zinc finger protein zfp47'

Table 2 (*Continued*)

99	'nuclear fragile X mental retardation protein interacting protein 1'
100	'ESTs, Highly similar to beta-1,3-*N*-acetylglucosaminyltransferase [*H. sapiens*]'
100 genes with maximum residuals	
101	'heterochromatin-like protein 1'
102	'SRY (sex determining region Y)-box 2'
103	'gap junction protein, alpha 1, 43 kDa (connexin 43)'
104	'nucleotide binding protein'
105	'COP9 homolog'
106	'ESTs, weakly similar to ALU1_human ALU subfamily j sequence contamination warning entry [*H. sapiens*]'
107	'serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 3'
108	'small nuclear ribonucleoprotein polypeptide N'
109	'CGI-136 protein'
110	'adenylate cyclase activating polypeptide 1 (pituitary)'
111	'EST'
112	'ESTs'
113	'ESTs'
114	'elastase 3B'
115	'UDP-*N*-acetyl-alpha-D-galactosamine:polypeptide *N*-acetylgalactosaminyltransferase 1 (GalNAc-T1)'
116	'ESTs, Weakly similar to AF132972 1 CGI-38 protein [*H. sapiens*]'
117	'DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 8 (RNA helicase)'
118	'ESTs'
119	'KIAA0123 protein'
120	'ectodermal-neural cortex (with BTB-like domain)'
121	'retinoid X receptor, beta'
122	'KIAA0381 protein'
123	3381
124	'homogentisate 1,2-dioxygenase (homogentisate oxidase)'
125	'macrophage stimulating 1 receptor (c-met-related tyrosine kinase)'
126	'amiloride binding protein 1 (amine oxidase (copper-containing))'
127	'jumping translocation breakpoint'
128	'ketohexokinase (fructokinase)'
129	7656
130	'protein tyrosine phosphatase, non-receptor type 1'
131	'CD83 antigen (activated B lymphocytes, immunoglobulin superfamily)'
132	'developmentally regulated GTP-binding protein 2'
133	'ubiquitin carrier protein'
134	'RNA-binding protein (autoantigenic)'
135	'membrane protein, palmitoylated 2 (MAGUK p55 subfamily member 2)'
136	'runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene)'
137	'small nuclear ribonucleoprotein polypeptides B and B1'
138	204
139	'RD RNA-binding protein'
140	'somatostatin'
141	'protein phosphatase 2 (formerly 2A), regulatory subunit B (PR 52), beta isoform'
142	'crystallin, gamma C'
143	'ligase III, DNA, ATP-dependent'
144	'actin binding LIM protein 1'
145	'protein kinase C binding protein 2'
146	'ESTs'
147	'nucleoporin 214 kDa (CAIN)'
148	'ESTs'
149	356
150	'TAP binding protein (tapasin)'
151	'eyes absent (*Drosophila*) homolog 3'
152	'HNK-1 sulfotransferase'
153	'ESTs'
154	'keratin 6A'
155	'hypothetical protein FLJ20371'

Table 2 (*Continued*)

156	'interferon-related developmental regulator 2'
157	'tyrosyl-tRNA synthetase'
158	'SCAN domain-containing 1'
159	7787
160	'GTP cyclohydrolase 1 (dopa-responsive dystonia)'
161	'sodium-dependent high-affinity dicarboxylate transporter 3'
162	'small inducible cytokine subfamily B (Cys-X-Cys), member 6 (granulocyte chemotactic protein 2)'
163	'RaP2 interacting protein 8'
164	'pleckstrin homology-like domain, family A, member 1'
165	566
166	'pilin-like transcription factor'
167	'hypothetical protein FLJ10193'
168	'KIAA0769 gene product'
169	'*Homo sapiens* OSBP-related protein 6 mRNA, complete cds'
170	'CD47 antigen (Rh-related antigen, integrin-associated signal transducer)'
171	'*Homo sapiens* complement-c1q tumor necrosis factor-related protein (CTRP2) mRNA, complete cds'
172	3814
173	'copine VI (neuronal)'
174	'DKFZP586G1122 protein'
175	'pleckstrin homology, Sec7 and coiled/coil domains 4'
176	'zinc finger protein 219'
177	'ESTs'
178	'echinoderm microtubule-associated protein-like'
179	'hypothetical protein MGC5350'
180	5430
181	'transcription factor Dp-2 (E2F dimerization partner 2)'
182	'heterogeneous nuclear ribonucleoprotein R'
183	'KIAA0346 protein'
184	'ESTs'
185	'phosphoglycerate kinase 1'
186	'heterogeneous nuclear ribonucleoprotein A1'
187	'hypothetical protein DKFZp434B217'
188	'exostoses (multiple)-like 2'
189	'hypoxanthine phosphoribosyltransferase 1 (Lesch-Nyhan syndrome)'
190	'SRY (sex determining region Y)-box 3'
191	'retinoblastoma-binding protein 2'
192	'proteasome (prosome, macropain) subunit, beta type, 8 (large multifunctional protease 7)'
193	'ZW10 (*Drosophila*) homolog, centromere/kinetochore protein'
194	6587
195	'inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase complex-associated protein'
196	'putative mitochondrial outer membrane protein import receptor'
197	'NADH dehydrogenase (ubiquinone) Fe—S protein 6 (13 kDa) (NADH-coenzyme Q reductase)'
198	'LPAP for lysophosphatidic acid phosphatase'
199	'MyoD family inhibitor'
200	'sperm associated antigen 11'

parameter set in the consequent parts. Let e_{ij} be the error between the jth desired output and the output of the ith rule with the jth input data; i.e.

$$e_{ij} = y_j - f_i(\vec{x}(j); \vec{a}^i), \quad i = 1, 2, \ldots, C \quad \text{and} \quad j = 1, 2, \ldots, N \tag{15}$$

The fitness function is defined as

$$J = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^2 e_{ij}^2 - \sum_{i=1}^{C}\left(\sum_{j=1}^{N} u_{ij}\right)^2 \tag{16}$$

subject to

$$\sum_{i=1}^{C} u_{ij} = 1, \quad 1 \le j \le N \tag{17}$$

where u_{ij} is the firing strength of the ith rule for the jth training pattern. To minimize fitness function J in Eq. (16) subject to Eq. (17), it is easy to show that the parameter vector $\vec{a}^i$ for the consequent part of the ith rule can be obtained as [50]

$$\vec{a}^i = [X^{\mathrm{T}} D_i X]^{-1} X^{\mathrm{T}} D_i Y, \quad i = 1, 2, \ldots, C \tag{18}$$

where $X \in R^{N\times(n+1)}$ is the matrix with $\vec{x}(k)$ as its $(k+1)$th row and the elements in the first row are all 1, $Y \in R^N$ a vector with y_k as its kth element, and $D_i \in \mathrm{R}^{N\times N}$ is a diagonal matrix with u_{ik}^2 as its kth diagonal element. Then, u_{ij} is computed as follows [40,51]:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}(e_{ij}/e_{kj})^2} \tag{19}$$

Assume that Gaussian membership functions are used in the premise parts (i.e. $A_j^i(\Omega_j^i, x_j(i)) = \exp\{-(x_j(i)-\Omega_{j1}^i)^2/2(\Omega_{j2}^i)^2\}$), where Ω_{j1}^i and Ω_{j2}^i are parameters of the jth membership function for the ith fuzzy rules. Then, they can easily be computed as follows:

$$\Omega_{j1}^i = \frac{\sum_{k=1}^{N}(u_{ik})^2 x_j(k)}{\sum_{k=1}^{N}(u_{ik})^2} \tag{20}$$

$$\Omega_{j2}^i = \sqrt{\frac{\sum_{k=1}^{N}(u_{ik})^2 (x_j(k)-\Omega_{j1}^i)^2}{\sum_{k=1}^{N}(u_{ik})^2}} \tag{21}$$

Finally, the proposed algorithm is repeated until the termination criterion has satisfied.

4. Simulation results

In simulations, we need to identify a set of values. The size of the initial population for GA and particles of PSO are both set as 20. The crossover probability $P_c = 0.7$, mutation probability $P_m = 0.02$, MAXGEN = 100, $c_1 = 1$, $c_2 = 1$, $\beta_1 = 0.9$, $\beta_2 = 0.4$, $l = 10$, $C \in [1\ 1000]$, $\sigma \in [1\ 100]$, $W_a = 5000$, $W_b = 10$, and *run_num* = 200.

These 200 target genes obtained after regression analysis is shown in Table 2. The cardinalities of selected target genes are calculated by the hybrid process of GA, PSO and SVM. Six gene markers are extracted after ANOVA. These genes are 'complement component 1, s subcomponent', 'Human insulin-like growth factor binding protein 5 (IGFBP5) mRNA', '*Homo sapiens* mRNA; cDNA DKFZp434A115 (from clone DKFZp434A115)', 'GL002 protein', '7656', 'ZW10 (*Drosophila*) homolog, centromere/kinetochore protein'. After obtaining six gene markers, the improved fuzzy model and some commonly used methods such as fuzzy *c*-means (FCM), hierarchical clustering, *k*-means and linear discriminant analysis (LDA) are used to verify OVT and OVCA. FCM, hierarchical clustering, and *k*-means are famous clustering techniques [40,41]. LDA is an analysis of dependence method that is a special case of canonical correlation [47]. The simulation results are tabulated in Tables 3–5 and Fig. 3. From Tables 3–5, the results for the improved fuzzy model and FCM are superior to *k*-means and LDA. Both methods of *k*-means and LDA cannot successively classify OVT and OVCA. From Table 3, it shows that improved fuzzy model and FCM can successfully classify five OVT and five OVCAI, but one OVCAIII is misclassified as OVCAI (not in the same cluster). The hierarchical clustering also has the same result in Fig. 3. The pathology of this OVCAIII is "Mucinous cytadenoma, borderline malignancy" and the patient's age is 26. Because this sample is the youngest woman among all patients in Table 1, it might be defined as a new subtype (OVTT) instead of OVCAIII [44]. The improved fuzzy model, FCM, and hierarchical clustering are performed to verify this sample, and the simulation results are listed in Tables 6 and 7, and Fig. 4. From Table 7, only the improved fuzzy model can successively classify OVT, OVTT, OVCAI

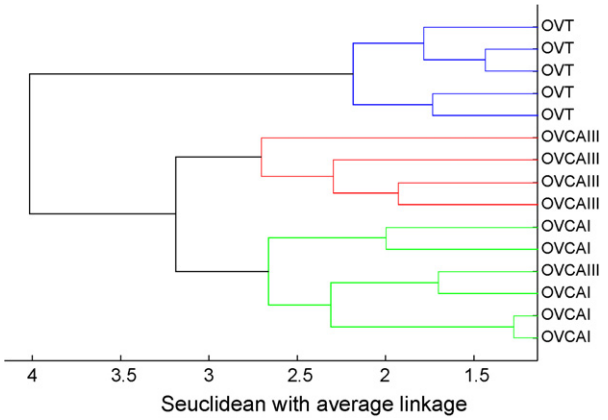

Figure 3 The result of hierarchical clustering for OVT, OVCAI and OVCAIII.

Table 3 The results of the improved fuzzy method and FCM for OVT, OVCAI and OVCAIII

Cluster no.	OVT	OVCA		Total
		OVCAI	OVCAIII	
1	5	0	0	5
2	0	5	1	6
3	0	0	4	4
Total	5	5	5	15

Table 4 The results of *k*-means for OVT, OVCAI and OVCAIII

Cluster no.	OVT	OVCA		Total
		OVCAI	OVCAIII	
1	4	1	1	6
2	1	4	1	6
3	0	0	3	3
Total	5	5	5	15

Table 5 The results of LDA for OVT, OVCAI and OVCAIII

Classification no.	OVT	OVCA		Total
		OVCAI	OVCAIII	
1	4	1	0	5
2	1	3	2	6
3	0	1	3	4
Total	5	5	5	15

Table 6 The results of FCM for OVT, OVTT, OVCAI and OVCAIII

Cluster no.	OVT	OVCA			Total
		OVTT	OVCAI	OVCAIII	
1	4	0	1	0	5
2	0	0	4	0	4
3	0	0	0	4	4
4	1	1	0	0	2
Total	5	1	5	4	15

Table 7 The results of the improved fuzzy method for OVT, OVTT, OVCAI and OVCAIII

Cluster no.	OVT	OVCA			Total
		OVTT	OVCAI	OVCAIII	
1	5	0	0	0	5
2	0	0	5	0	5
3	0	0	0	4	4
4	0	1	0	0	1
Total	5	1	5	4	15

and OVCAIII, but FCM and hierarchical clustering cannot successively classify these subtypes. It indeed shows the superior performance for the proposed algorithm. Finally, two data sets are also used to test the performance of the proposed algorithm. The first data consists of expression profiles of 2000 genes from 22 normal and 40 colon tumor tissues which are taken from http://microarray.-princeton.edu/oncology/affydata/index.html (accessed: 15 September 2007) [52]. The second data set consists of breast cancer form 683 samples which are taken from UCI machine learning repository [53]. The hybrid process of SVM and GA, the hybrid process of SVM and PSO, and the proposed algorithm are used to evaluate the performance of obtained gene markers. In simulation, 70% of samples are used as training data and 30% samples are randomly selected to test the accuracy of classification. It is noted that the parameters of all compared approaches are set as the same values of the proposed algorithm, and the accuracy of classification is verified by the improved fuzzy model. The simulation results are listed in Table 8. It is easy to see that proposed algorithm has the best accuracy among those approaches. The proposed algorithm outperforms other approaches, as expected.

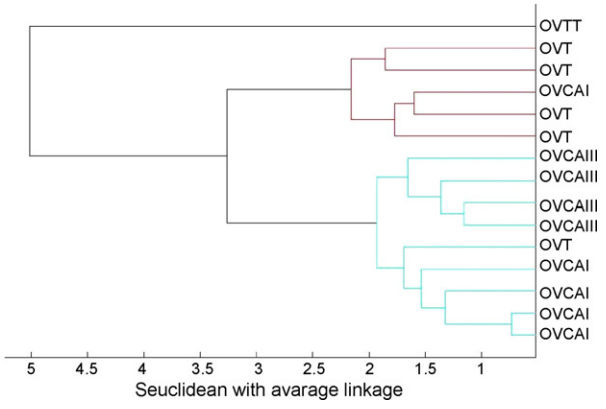

Figure 4 The result of hierarchical clustering for OVT, OVTT, OVCAI and OVCAIII.

Table 8 The accuracy of classification for various approaches

	The hybrid process of SVM and GA (%)	The hybrid process of SVM and PSO (%)	The proposed algorithm (%)
Colon	95.65	97.13	99.13
Breast	96.23	97.95	98.55

5. Conclusions

For cancer research, it is difficult to search the sensitive and specific gene markers. In this paper, the proposed algorithm can analyze gene expressions for microarray data of ovarian cancer. The proposed algorithm can find target genes after regression analysis. Additionally, the process of GA, PSO, SVM and ANOVA are hybridized to select gene markers. In this hybrid process, the parameters of C and σ for SVM are automatically achieved. Furthermore, the obtained gene markers are performed to classify cancer tissues by the improved fuzzy model. It has demonstrated that six gene markers can be obtained through the proposed algorithm. These gene markers can successively classify the subtype of OVT, OVTT, OVCAI and OVCAIII. Because the proposed algorithm has superior performance for gene selection and classification in microarray data of ovarian cancer, it can also be performed on other studies for cancer diagnosis. The basic results of this study can be giving as guidance for biologists in future work.

Acknowledgements

Author appreciates G. Steven Huang, Yao-Ching Hung, Meng-Hsiun Tsai, Yen-Po Huang, Shih-Chieh Chen, and Shih-Wei Lin for providing the microarray data and discussions, and thanks the anonymous reviewers for their comments and constructive suggestions that have improved this paper. This work was supported by National Science Council under grant NSC 96-2221-E-211-021.

References

[1] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467—70.
[2] Jemal A, Thomas A, Murray T, Thun M. Cancer statistics. CA: Cancer J Clin 2002;52:23—47.
[3] Steven Huang G, Hung Y-C, Chen A, Hong M-Y. Microarray analysis of ovarian cancer. In: IEEE international conference on systems, man and cybernetics. New York: IEEE Systems, Man and Cybernetics Society; 2005. p. 1036—42.
[4] Jermal A, Murray T, Samuels A, Ghafoor A, Ward E, Thun MJ. Cancer statistic 2003. CA: Cancer J Clin 2003;13:5—26.
[5] Knudsen S. A biologist's analysis of DNA microarray data. New York: John Wiley & Sons Inc., Publication; 2002.
[6] Jeng J-T, Lee T-T, Lee Y-C. Classification of ovarian cancer based on intelligent systems with microarray data. In: IEEE international conference on systems, man and cybernetics. New York: IEEE Systems, Man and Cybernetics Society; 2005 p. 1053—8.
[7] Chuang C-C, Jeng J-T, Su S-F. Dimension reduction with support vector regression for ovarian cancer microarray data. In: IEEE international conference on systems, man and cybernetics. New York: IEEE Systems, Man and Cybernetics Society; 2005. p. 1048—52.
[8] Chao S, Lihui C. Feature dimension reduction for microarray data analysis using locally linear embedding. In: Phoebe Chen Y-P, Wong L, editors. Proceeding of 3rd Asia-Pacific bioinformatics conference. London: Imperial College Press; 2005. p. 211—7.
[9] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503—11.
[10] DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457—60.
[11] Chen D, Hua D, Reifman J, Cheng X. Gene selection for multiclass prediction of microarray data. In: Proceedings of the IEEE computer society conference on bioinformatics. Washington: IEEE Computer Society; 2003. p. 492—5.
[12] Lee KE, Sha N, Dougherty E, Vannucci M, Mallick BK. Gene selection: a Bayesian variable selection approach. Bioinformatics 2003;19:90—7.
[13] Park PJ, Pagano M, Bonetti M. A nonparametric scoring algorithm for identifying informative genes from microarray data. In: Altman RB, Keith Dunker A, Hunter L, Lauderdale K, Klein TE, editors. Pacific symposium on biocomputing 2001. New Jersey: World Scientific; 2001. p. 52—63.
[14] Tan TZ, Quek C, Ng GS. Ovarian cancer diagnosis using complementary learning fuzzy neural network. In: Proceedings of the 2005 IEEE international joint conference on neural networks (IJCNN 2005). Piscataway, New Jersey: IEEE Service Center; 2005. p. 3034—9.
[15] Schaffer JD, Janevski A, Simpson MR. A Genetic Algorithm approach for discovering diagnostic patterns in molecular measurement data. In: Proceedings of the 2005 IEEE symposium on computational intelligence in bioinformatics and computational biology. La Jolla, California: IEEE Computer Society; 2005. p. 1—8.
[16] Bertoni A, Valentini G. Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses. Artif Intell Med 2006;37:85—109.
[17] Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7:673—9.
[18] Chu F, Xie W, Wang L. Gene selection and cancer classification using a fuzzy neural network. In: Dick S, Kurgan L, Musilek P, Pedrycz W, Reformat M, editors. Proceedings of NAFIPS 2004, Annual meeting of the North American fuzzy information processing society. Banff, Alberta, Canada: IEEE; 2004. p. 555—9.
[19] Ni B, Li J. A hybrid filter/wrapper gene selection method for microarray classification. In: Proceedings of 2004 international conference on machine learning and cybernetics, vol. 4. New York: IEEE Systems, Man and Cybernetics Society; 2004. p. 2537—42.
[20] Hong J-H, Cho S-B. The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming. Artif Intell Med 2006;36:43—58.
[21] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16:906—14.
[22] Huang TM, Kecman V. Gene extraction for cancer diagnosis by support vector machines—an improvement. Artif Intell Med 2005;35:185—94.
[23] Ma PCH, Keith C, Chan C, Yao X, Chiu DKY. An evolutionary clustering algorithm for gene expression microarray data analysis. IEEE Trans Evol Comput 2006;10:296—314.
[24] Schummer M, Ng WV, Bumgarner RE, Nelson PS, Schummer B, Bednarski DW, et al. Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene 1999;85:238—375.
[25] Wang K, Gan L, Jeffery E, Gayle M, Gown AM, Skelly M, et al. Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. Gene 1999;229:101—9.
[26] Ismail RS, Baldwin RL, Fang J, Browning D, Karlan BY, Gasson JC, et al. Differential gene expression between normal and tumor-derived ovarian epithelial cells. Cancer Res 2000;60:6744—53.
[27] Martoglio AM, Tom BD, Starkey M, Corps AN, Charnock-Jones DS, Smith SK. Changes in, tumor-genesis and angiogenesis-related gene transcript, abundance profiles in ovarian cancer detected by tailored high density cDNA arrays. Mol Med 2000;6:750—65.
[28] Ono K, Tanaka T, Tsunoda T, Kitahara O, Kihara C, Okamoto A, et al. Identification by cDNA microarray of genes involved in ovarian carcinogenesis. Cancer Res 2000;60:5007—18.
[29] Welsh JB, Zarrinkar PP, Sapinoso LM, Kern SG, Behling CA, Monk BJ, et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. In: Proceedings of the national academy of sciences of the United States of America, vol. 98. Washington: National Academy of Sciences; 2001. p. 1176—81.
[30] Gen M, Cheng R. Genetic algorithms and engineering design. New York: John Wiley & Sons Inc., Publication; 1997.

[31] Lee Z-J, Su S-F, Lee C-Y. Efficiently solving general weapon-target assignment problem by genetic algorithms with greedy eugenics. IEEE Trans Syst Man Cybernetics Part B 2003;33:113—21.
[32] Lee Z-J, Su S-F, Lee C-Y, Hung Y-S. A heuristic genetic algorithm (HGA) for solving resource allocation problems. Knowledge Inform Syst 2003;5:503—11.
[33] Lee Z-J, Wang Y-P, Su S-F. A Genetic algorithm based robust learning credit assignment cerebellar model articulation controller. Appl Soft Comput 2004;4:357—67.
[34] Kennedy J, Eberhart R. Particle swarm optimization. In: Proceedings of IEEE international conference on neural networks. Piscataway, New Jersey: IEEE Service Center; 1995 p. 1942—8.
[35] Coello CAC, Pulido GT, Lechuga MS. Handling multiple objectives with particle swarm optimization. IEEE Trans Evol Comput 2004;8:256—79.
[36] Huang C-M, Huang C-J, Wang M-L. A particle swarm optimization to identifying the ARMAX model for short-term load forecasting. IEEE Trans Power Syst 2005;20:1126—33.
[37] Weijun X, Wu Z. An effective hybrid optimization approach for multi-objective flexible job-shop scheduling problems. Comput Ind Eng 2005;48:409—25.
[38] Liang JJ, Qin AK, Suganthan PN, Baskar S. Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. IEEE Trans Evol Comput 2006;10: 281—95.
[39] Vapnik VN. The nature of statistical learning theory. Berlin: Springer-Verlag; 1995.
[40] Gomez-Skarmete AF, Delgado M, Vila MA. About the use of fuzzy clustering techniques for fuzzy model identification. Fuzzy Sets Syst 1999;106:179—88.
[41] Jain AK, Dubes RC. Algorithms for clustering data. New Jersey: Prentice Hall; 1988.
[42] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29:365—71.
[43] Eisen M. ScanAlyze user manual. USA: Standford University; 1999.
[44] Lu S-J. Gene expression analysis and regulator pathway exploration with the use of microarray data for ovarian cancer. Master thesis. National Taiwan University of Science and Technology, Taiwan; 2006.
[45] Rossi F, Villa N. Support vector machine for functional data classification. Neurocomputing 2006;69:730—42.
[46] Hogg RV, Ledolter J. Engineer statistics. England: MacMillan Publishing Company; 1987.
[47] Hand D, Mannie H, Smyth P. Principles of data mining. Cambridge, Massachusetts: The MIT Press; 2001.
[48] Juang C-F. A hybrid of genetic algorithm and particle swarm optimization for recurrent network design. IEEE Trans Syst Man Cybernetics Part B 2004;34:997—1006.
[49] Esmin AAA, Lambert-Torres G, Alvarenga GB. Hybrid evolutionary algorithm based on PSO and GA mutation. In: Proceedings of the sixth international conference on hybrid intelligent systems (HIS'06). Los Alamitos, California: IEEE Computer Society; 2006. p. 57—63.
[50] Lee Z-J. A novel hybrid algorithm for function approximation. Expert Syst Appl 2008;34:384—90.
[51] Dave RN, Krishnapurum R. Robust clustering methods: a unified view. IEEE Trans Fuzzy Syst 1997;5:270—93.
[52] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In: Proceedings of the national academy of sciences of the United States of America, vol. 96. Washington: National Academy of Sciences; 1999 p. 6745—50.
[53] Hettich S, Blake C, Merz C. UCI repository of machine information and computer sciences. Available at http://www.ics.uci.edu/~mlearn/MLRepository.html (accessed: September 15, 2007).