Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 1975 – 1984
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

A novel approach to feature extraction from classification models based on information gene pairs

J. Li[*], X. Tang, J. Liu, J. Huang, Y. Wang

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

Received 11 November 2006; received in revised form 19 September 2007; accepted 20 November 2007

Abstract

Various microarray experiments are now done in many laboratories, resulting in the rapid accumulation of microarray data in public repositories. One of the major challenges of analyzing microarray data is how to extract and select efficient features from it for accurate cancer classification. Here we introduce a new feature extraction and selection method based on information gene pairs that have significant change in different tissue samples. Experimental results on five public microarray data sets demonstrate that the feature subset selected by the proposed method performs well and achieves higher classification accuracy on several classifiers. We perform extensive experimental comparison of the features selected by the proposed method and features selected by other methods using different evaluation methods and classifiers. The results confirm that the proposed method performs as well as other methods on acute lymphoblastic-acute myeloid leukemia, adenocarcinoma and breast cancer data sets using a fewer information genes and leads to significant improvement of classification accuracy on colon and diffuse large B cell lymphoma cancer data sets.

Keywords: Feature extraction; Information gene pair; Microarray data; Cancer classification; Genetic algorithm

1. Introduction

The introduction of DNA microarray technology has made it possible to acquire vast amounts of microarray data, raising the issue of how best to extract and select features from this data. Various methods have been proposed for extracting and selecting features from microarray data. Principal component analysis (PCA) is widely used to analyze image and speech data. It is also used to extract features from microarray data [1,2]. However, the features obtained through the method are short of clear biological meaning and cannot help biologists find important information genes [2]. This is especially important in microarray data-based cancer classification. At present, the most commonly used methods for selecting feature genes from microarray data are filter methods that rank genes according to some predefined criterion and select the top-ranked genes. For example, t-test [3], signal-noise-rate (SNR) [4,5] and Wilcoxon's ranksum test (WRST) [6] are typical filter methods. Filter methods are easy to understand and implement; however, they ignore the interrelation of genes, this is inevitable to lose some important information. In addition, the classification accuracy of the feature genes selected by filter methods may be lower because the top-ranked k genes are not guaranteed to be the best among all subsets of k genes. Wrapper methods have also been widely used to select feature genes from microarray data [7–11]. They evaluate alternative feature gene subsets using the classification accuracy and select the feature gene subset with the highest classification accuracy. Compared with the feature genes selected by filter methods, the feature genes selected by wrapper methods usually have higher classification accuracy.

Although various methods have been used to extract and select features from microarray data, development of powerful and efficient feature extraction and selection approach to improve the performance of cancer classification remains a significant demand. In this study, we proposed a novel feature extraction method that treats information gene pair, which is highly correlative in one type of tissue sample and has significant change in another type of tissue sample, as atomic unit

[*] Corresponding author.
E-mail address: jieleehz@yahoo.com.cn (J. Li).

0031-3203/$30.00 © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2007.11.019

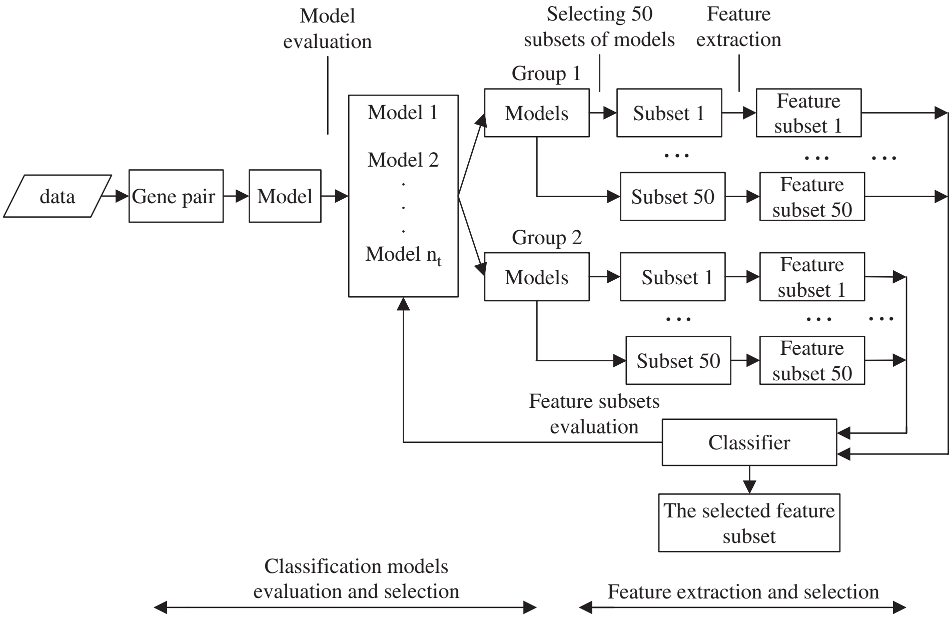

Fig. 1. The main processes involved in the proposed method.

extracts features from the classification models based on the information gene pairs. Fig. 1 shows the main steps that the proposed method extracts features. First, we construct the classification models based on information gene pairs and evaluate their performance using classification accuracy. The top-ranked n_t classification models with higher classification accuracy are input to the next step. Then, the n_t classification models are divided into two groups. Group 1 consists of classification models based on information gene pairs that are highly correlative in class 1. Group 2 consists of classification models based on information gene pairs that are highly correlative in class 2. Subsequently, we use genetic algorithm (GA) to select an optimal subset of classification models from groups 1 and 2, respectively, and extract feature subset from them. Finally, the feature subset with the best performance is selected.

The rest of this paper is organized as follows. The method of extracting features from microarray data is elaborated in Section 2. In Section 3, we test our method on several microarray data sets and compare the performance of the features selected by the proposed method with that of the features selected by other methods. Finally conclusions are given.

2. Method

Definition. Different tissue samples, such as cancer and normal tissue samples or liver cancer and non liver cancer tissue samples (both of which are cancer tissue samples), often are examined in a microarray experiment. For two genes: g1 and g2, examined in two types of tissue samples (for example, normal and cancer tissue samples), when they have the following characteristics:

- They are highly correlative in class 1 (or class 2).
- The expression levels of g1 or/and g2 have significant changes that make two types of samples separable

They are called information gene pair.

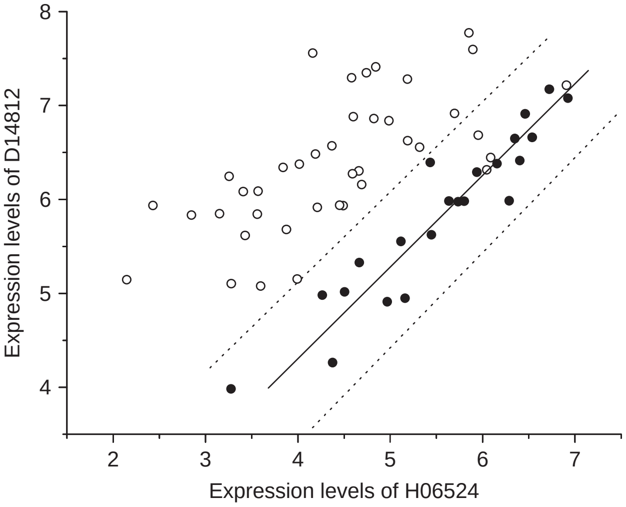

Fig. 2. The expression levels of information gene pair: H06524 and D14812. The *x*-axis represents the expression level of H06524, and the *y*-axis represents the expression level of D14812. The points marked "●" are normal samples, and the cancer samples are marked by "○". The data set has been preprocessed by taking logarithm of all values.

Figs. 2 and 3 show the distribution of expression values of two typical information gene pairs. D14812 and H06524 are a pair of information genes from colon cancer data [11], which are highly correlative in the normal samples (correlation coefficient is 0.93). The expression levels of D14812 in colon cancer samples are higher and the expression levels of H06524 in colon cancer samples are lower, which make the colon cancer samples and normal samples separable. Fig. 2 clearly shows the above characteristics. D63874 and R59552 are another pair of information genes from adenocarcinoma cancer data [12]. From Fig. 3 (classification model based on the pair of genes), it can be seen that the pair of genes is highly correlative in

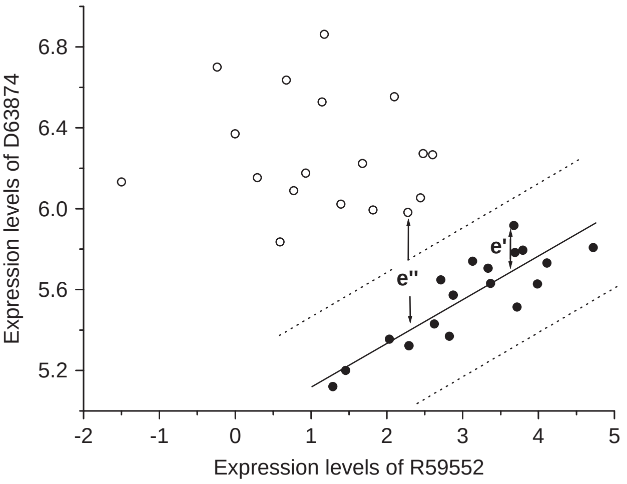

Fig. 3. The expression levels of information gene pair: R59552 and D63874. The *x*-axis represents the expression level of R59552, and the *y*-axis represents the expression level of D63874. The points marked "●" are normal samples, and the cancer samples are marked by "○". The data set has been preprocessed by taking logarithm of all values.

normal samples (correlation coefficient is 0.86) and the expression levels of D63874 have significant changes in normal and cancer samples. The classification model based on the pair of information genes gives 100% separation between normal and cancer samples.

2.1. Classification model based on information gene pair

Suppose two types of tissue samples are examined in a microarray experiment, k is the number of genes, $n1$ and $n2$ ($n = n1 + n2$) are the number of samples in classes 1 and 2, respectively, we can describe the microarray data using two matrices: $Y = (y_{ip})_{k\times n}$, $X = (x_{iq})_{k\times n}$, where $y_{ip}(x_{iq})$ denotes the expression level of the ith gene in the pth (qth) sample which belongs to class 1 (2). Given the ith gene and the jth gene (a pair of information genes) are highly correlative in class 1, for the pth sample from class 1, we can predict y_{ip} via the following regression model:

$$\widehat{y}_{ijp} = \widehat{\beta}_{ij0} + \widehat{\beta}_{ij1} y_{jp}, \quad 1 \leqslant p \leqslant n1. \tag{1}$$

$\widehat{\beta}_{ij0}$ and $\widehat{\beta}_{ij1}$ are estimated from a set of data, (y_{i1}, y_{j1}), $(y_{i2}, y_{j2}), \ldots, (y_{in1}, y_{jn1})$, using the least squares methods.

Define residual value $e_{ijp} = |y_{ip} - \widehat{y}_{ijp}|$ as the difference between the observed value y_{ip} and the predicted value $\widehat{y}_{ijp}$. For all the samples from class 1, we have

$$E1_{ij} = \{e_{ijp} | e_{ijp} = |y_{ip} - \widehat{\beta}_{ij0} - \widehat{\beta}_{ij1} y_{jp}|, 1 \leqslant p \leqslant n1\}. \tag{2}$$

For the qth sample from class 2, we still use the model (1) to predict x_{iq}, then the predicted value is

$$\widehat{x}_{ijq} = \widehat{\beta}_{ij0} + \widehat{\beta}_{ij1} x_{jq}, \quad 1 \leqslant q \leqslant n2.$$

The residual value is $e_{ijq} = |x_{iq} - \widehat{x}_{ijq}|$.

For all the samples from class 2, we have

$$E2_{ij} = \{e_{ijq} | e_{ijq} = |x_{iq} - \widehat{\beta}_{ij0} - \widehat{\beta}_{ij1} x_{jq}|, 1 \leqslant q \leqslant n2\}. \tag{3}$$

Here, model (1) projects the expression values of the ith gene in two types of samples into two subsets: $E1_{ij}$ and $E2_{ij}$. The problem of discriminating two types of samples has become the problem of discriminating the elements in two subsets. In the following, we give the classification rule of regression model according to the optimal threshold value e_d which minimizes the error of discriminating the elements in two subsets: $E1_{ij}$ and $E2_{ij}$.

Define function

$$f_i(e) = count(\{e_{ijp} < e, e_{ijp} \in E1_{ij}, n1 \geqslant p \geqslant 1\} \cup \{e_{ijq} > e, e_{ijq} \in E2_{ij}, n2 \geqslant q \geqslant 1\}), \tag{4}$$

where e is a real number, count $(\cdot)$ denotes the number of elements in the subset. Let $e = e_1$, $f_i(e_1) = \max(f_i(e))$, the threshold value e_d is

$$e_d = (\max(\{e_{ijp} | e_{ijp} \leqslant e_1, e_{ijp} \in E1_{ij}\}) + \min(\{e_{ijq} | e_{ijq} > e_1, e_{ijq} \in E2_{ij}\}))/2. \tag{5}$$

Thus, selecting a sample randomly from the total samples, the expression levels of the ith and the jth gene in the sample are w_i and w_j, respectively, classification can be achieved according to the following rule:

Assign the sample in class 1 if $|w_i - \widehat{w}_i| \leqslant e_d$, namely, $|w_i - \widehat{\beta}_{ij0} - \widehat{\beta}_{ij1} w_j| \leqslant e_d$, and in class 2 otherwise.

Fig. 3 provides a geometric interpretation of the proposed classification model. In Fig. 3,

$$e' = \max(\{e_{ijp} | e_{ijp} < e_1, e_{ijp} \in E1_{ij}\}),$$

$$e'' = \min(\{e_{ijq} | e_{ijq} < e_1, e_{ijq} \in E2_{ij}\}),$$

$$e_d = (e' + e'')/2,$$

solid line is a regression line, two dash lines which are parallel to the regression line are decision boundary. The samples between two dot lines are assigned to normal samples (class 1), the others are assigned to adenocarcinoma cancer samples (class 2). Classification model based on D63874 and R59552 can classify adenocarcinoma cancer samples and normal samples 100%.

2.2. Classification model evaluation and selection

There exist a large number of the proposed classification models in a microarray data; however, many of them could be either redundant or even irrelevant to the classification task; thus, we need to filter out those redundant or irrelevant classification models and select the classification models with better classification performance.

There are three methods that are widely used in evaluating the performance of classification model [9]. When the collection of total samples are used as both training and test data sets, the classification accuracy is referred to as the within sample classification accuracy (WSCA). When the training and test samples are separate data sets, the classification accuracy is referred to the out-of-sample classification accuracy because test samples are used for the calculation of accuracy. Bootstrapping method is developed for overcoming the problems of small data set and better assessing the performance of classification model. The classification accuracy obtained through the method is called bootstrapping accuracy. Compared with the above two methods, bootstrapping method needs more computational time.

In the microarray data that includes k genes and $n(n = n1 + n2)$ samples ($n1$ and $n2$ are the number of samples in classes 1 and 2, respectively), there exist $2 \times n \times (n-1)$ classification models. It needs too much time to evaluate all classification models. To reduce computational time and obtain the classification models with better classification ability, the classification models based on gene pairs whose correlation coefficients are lower than threshold ∂_h in classes 1 and 2 are not evaluated. The others are evaluated using a two-step procedure. (1) Each classification model is first evaluated using WSCA. The top-ranked n_t ones are input to the next phase. (2) Bootstrapping procedure is used to further evaluate the performance of the selected n_t classification models. There are a number of variants of bootstrapping method. Here we used a straightforward one (see Ref. [13]). The n_t classification models are sorted in decreasing order based on bootstrapping accuracy and divided into two groups. Group 1 consists of classification models based on information gene pairs that are highly correlative in class 1. Group 2 includes the classification models based on information gene pairs that are highly correlative in class 2. We extract feature subsets from the top-ranked classification models in groups 1 and 2, respectively.

2.3. Feature extraction from classification models

In Section 2.1, classification model based on the information gene pair (the ith gene and the jth gene) projects the expression values of the ith gene in two types of samples into two subsets:$E1_i$ and $E2_i$. For $m1$ pairs of information genes, $(i_1, j_1), \ldots, (i_{m1}, j_{m1})$, which are highly correlative in class 1, we can construct $m1$ linear regression models that project the expression values of the $m1$ genes $(i_1, i_2, \ldots, i_{m1})$ in two types of samples into $m1$ pairs of subsets:

$$(E1_{i_1 j_1}, E2_{i_1 j_1}), (E1_{i_2 j_2}, E2_{i_2 j_2}), \ldots, (E1_{i_{m1} j_{m1}}, E2_{i_{m1} j_{m1}}).$$

For the pth sample from class 1 and the qth sample from class 2, we have

$$\mu 1_p = \frac{1}{m1} \sum_{l=1}^{m1} e_{i_l j_l p}, \quad e_{i_l j_l p} \in E1_{i_l j_l}, \tag{6}$$

$$\mu 2_q = \frac{1}{m1} \sum_{l=1}^{m1} e_{i_l j_l q}, \quad e_{i_l j_l q} \in E2_{i_l j_l}. \tag{7}$$

Here we choose $\mu 1_p$ as the feature of the pth sample and $\mu 2_q$ as the feature of the qth sample. For all samples in microarray data, we have feature subset:

$$U = \{\mu 1_1, \mu 1_2, \ldots, \mu 1_{n1}, \mu 2_1, \mu 2_2, \ldots, \mu 2_{n2}\}. \tag{8}$$

Similarly, for the $m2$ pairs of information genes, $(i'_1, j'_2), \ldots, (i'_{m2}, j'_{m2})$, which are highly correlative in class 2, $m2$ linear regression models based on the $m2$ pairs of genes project the expression values of the $m2$ genes $(i'_1, i'_2, \ldots, i'_{m2})$ in two types of samples into $m2$ pairs of subsets:

$$(E1_{i'_1 j'_1}, E2_{i'_1 j'_1}), (E1_{i'_2 j'_2}, E2_{i'_2 j'_2}), \ldots, (E1_{i'_{m2} j'_{m2}}, E2_{i'_{m2} j'_{m2}}).$$

For the pth sample from class 1 and the qth sample from class 2, we have

$$\mu 1'_p = \frac{1}{m2} \sum_{l=1}^{m2} e_{i'_l j'_l p}, \quad e_{i'_l j'_l p} \in E1_{i'_l j'_l}, \tag{9}$$

$$\mu 2'_q = \frac{1}{m2} \sum_{l=1}^{m2} e_{i'_l j'_l q}, \quad e_{i'_l j'_l q} \in E2_{i'_l j'_l}. \tag{10}$$

Here we can also choose $\mu 1'_p$ as the feature of the pth sample and $\mu 2'_q$ as the feature of the qth sample. For all samples in microarray data, we have the second feature subset:

$$U' = \{\mu 1'_1, \mu 1'_2, \ldots, \mu 1'_{n1}, \mu 2'_1, \mu 2'_2, \ldots, \mu 2'_{n2}\}. \tag{11}$$

2.4. Feature selection

The classification performance of the features extracted from the top-ranked m classification models in group 1 (or 2) according to the method described in Section 2.3 is not guaranteed to be the best. Here we employ GA to find an optimal subset of classification models from the top-ranked classification models in group 1 or 2 and extract feature subset from them.

GA is an effective evolutionary optimization method [14]. It includes several components: chromosome (a mathematical entity, not the biological chromosomes), fitness function, selection operator, mutation operator and crossover operator. Chromosome is the core unit that has been used as a binary encoded representation of the solutions to the optimization problem. In order to use the GA, a set of chromosomes is first constructed to form a "population". Then, the population of chromosomes are evaluated according to the required fitness function and assigned a probability of survival proportional to their fitness. Subsequently, the GA manipulates the population of chromosomes using selection, crossover and mutation operators and passes them on to the next generation. The whole process is repeated until the population converges to a satisfactory solution or after a fixed number of generations.

GA is applied to our problem to find an optimal subset of classification models. Before GA is used to find an optimal subset of classification models, several parameters must be determined. In this study, the length of each chromosome is l, and each classification model occupies 1 bit. Value "1" or "0" of any bit means the present or absent of the corresponding classification model. The population size used is 50. The features extracted from the top-ranked classification models may obtain better classification accuracy at the beginning of GA, and therefore we initialize the population of chromosomes randomly using the top-ranked l classification models. To ensure that the best chromosomes are most likely to survive in the subsequent generation, the best two chromosomes are entered into the respective next generation, and the remaining 48 chromosomes are filled based on sampling that is weighted according to the relative fitness of the chromosomes in the parent generation (probabilistically). Relative fitness is

$$f_{ri} = f_i \bigg/ \sum_{k=1}^{48} f_k, \quad 1 \leqslant i \leqslant 48,$$

where f_i is the fitness value of ith chromosome. The probabilities of crossover and mutation are 0.9 and 0.05, respectively. Control parameters in these ranges have been proposed by several researchers to guarantee good performance on carefully chosen testbeds of objective functions [15]. Stopping criteria is the number of generations is larger than 200 and increase in optimal fitness value is lower than 0.0001 for 20 cycles.

The goal of finding an optimal subset of classification models is to extract feature subset that achieve the same or better classification performance using fewer genes. Therefore, we evaluate the performance of a subset of classification models using the performance of the feature subset extracted from them. The performances of feature subset contain three terms: (1) the classification accuracy of feature subset, (2) the margin of the classifier trained by feature subset and (3) the number of genes involved in feature subset. If feature subsets extracted from two subsets of classification models achieve the same classification accuracy, while the margin of the classifiers trained by them is different, the feature subset that can train classifier with larger margin is preferred. If two feature subsets have the same classification accuracy and margin, the subset with fewer genes is preferred. For the three terms, accuracy is our major concern. The next important term is the margin of classifier. To combine the three terms, we used the following fitness function:

$$fitness = \begin{cases} Acc + 10^{-4}\dfrac{(LC-Fn)}{LC} & \text{if } Acc < 1, \\ Acc + 10^{-2}\dfrac{Mg}{MM} + 10^{-4}\dfrac{(LC-Fn)}{LC} & \text{if } Acc = 1, \end{cases}$$

where Acc is the WSCA of the feature subset and Fn is the number of gene pairs involved in the features subset. LC is the length of chromosome. Mg/MM reflects the magnitude of classifier margin. The WSCA of feature subset is computed through the same classification rule of discriminating the subsets: $E1_{ij}$ and $E2_{ij}$ (see Section 2.1). For example, the WSCA of feature subset U (see formula (8): $U = \{\mu 1_1, \mu 1_2, \ldots, \mu 1_{n1}, \mu 2_1, \mu 2_2, \ldots, \mu 2_{n2}\}$, is computed using the following the rule:

Selecting a sample randomly from the total samples (its feature value is μ_i), assign the sample in class 1 if $|\mu_i| \leqslant \mu_d$ and in class 2 otherwise.

μ_d is the optimal threshold value that minimizes the error of discriminating the elements in two subsets: $\{\mu 1_1, \mu 1_2, \ldots, \mu 1_{n1}\}$, $\{\mu 2_1, \mu 2_2, \ldots, \mu 2_{n2}\}$.

If Acc is equal to 100%, Mg is given by

$$Mg = \min(\mu 2_1, \mu 2_2, \ldots, \mu 2_{n2}) - \max(\mu 1_1, \mu 1_2, \ldots, \mu 1_{n1}).$$

According to formula (8), MM is given by

$$MM = abs\left(\frac{1}{n1}\sum_{i=1}^{n1} \mu 1_i - \frac{1}{n2}\sum_{i=1}^{n2} \mu 2_i\right).$$

The accuracy term ranges roughly from 0.50 to 1. Mg/MM ranges from 0 to 1. The third term ranges from 0 to 0.0001. Based on the weights that we have assigned to each term, the accuracy term dominates the fitness value. This implies that, when the accuracy is lower than 100%, the individuals with higher accuracy will outweigh individuals with lower accuracy, no matter how many feature genes they contain, when the accuracy achieves 100%, the individuals with larger margin will outweigh individuals with smaller margin, no matter how many feature genes they contain.

The choosing of the weights of the three terms depends on many factors; we need to find the best balance between model compactness and classification performance. Under some cases, we prefer higher classification accuracy, no matter what the cost might be. If this is the case, the weight associated with the accuracy term should be very high. Under different situations, we might favor more compact models over accuracy, as long as the accuracy is within a satisfactory range. In this case, we should choose a higher weight for the third term.

3. Experiments

3.1. Data sets

We applied the proposed method to analyze five public microarray data sets: diffuse large B cell lymphoma (DLBCL) [10], colon cancer [11], adenocarcinoma cancer [12], acute lymphoblastic leukemia (ALL)/acute myeloid leukemia (AML) [16] and breast cancer data [17]. The DLBCL data consist of 4026 genes and 42 samples (21 germinal center B-like DLBCL and 21 actived B-like DLBCL). The data were originally filtered and log-transformed (base 2). The colon cancer data contain 62 tissue samples (22 normal tissue samples and 40 tumor tissue samples) and 2000 genes. The adenocarcinoma data set for discrimination between adenoma, adenocarcinoma and normal tissue samples consists of 7457 genes and 36 samples (18 normal tissue samples and 18 cancer tissue samples). ALL/AML data set has 12 582 genes and 52 samples (24 ALL, 28 AML). The breast data set, which was produced for the classification of BRCA1 mutation and others (7 BRCA1 mutation samples and 15 BRCA2 samples and sporadic samples), consists of 3226 genes and 22 samples. In the present work, the preprocessing and/or log transformation are performed on the data sets prior to analysis, this includes imputation of missing values in DLBCL data set (imputed by the KNN Impute algorithm [18] and log-transformation.

3.2. Experimental results

We performed a sequence of experiments in which we set the threshold value ∂_h to be 0.70, 0.75, 0.80 and 0.85, and evaluated classification models based on information gene pairs whose correlation coefficients are higher than ∂_h. Here we only report the experimental results ∂_h is 0.75 because the experimental results with different ∂_h values are similar.

Among all classification models based on gene pairs, there are 456 ones with 100% WSCA for ALL/AML cancer data, 146 for adenocarcinoma cancer data and 2322 for the breast cancer data. The reason that so many classification models with 100% classification accuracy are found in breast cancer data may be due to the small sample size in the breast data set. The number of the classification models whose WSCA is over 95% reaches to 1987 for ALL/AML, 1200 for adenocarcinoma, and 4000 for breast cancer data, respectively.

Because when the total samples are used to evaluate the accuracy of classification model, it is not clear that whether the two-gene model can obtain lower generalization error rate. To get a realistic estimate of classification accuracy, we used bootstrapping procedure to further examine the classification performance of the top-ranked n_t classification models. We randomly drew 200 bootstrapping sample sets from the total sample set and calculated the bootstrapping accuracy. As a result, there are five classification models with 100% bootstrapping accuracy to be identified for ALL/AML data set, eight for adenocarcinoma cancer data set, and three for breast cancer data set, respectively. The bootstrapping accuracy of the top-ranked 2000 classification models from each data set is shown in Fig. 4. Although the bootstrapping accuracy of classification models is somewhat lower than their WSCA, the bootstrapping accuracy of most classification models is still very high. The number of classification models whose bootstrapping accuracy is over 95% reaches to 1374, 448, 1350 for ALL/AML, adenocarcinoma, and breast cancer data set, respectively. That a large number of excellent classification models are found demonstrate the proposed method performs well in finding excellent classification models. This also provides more opportunity for biologists to choose disease diagnosis microarray according to the experiment conditions.

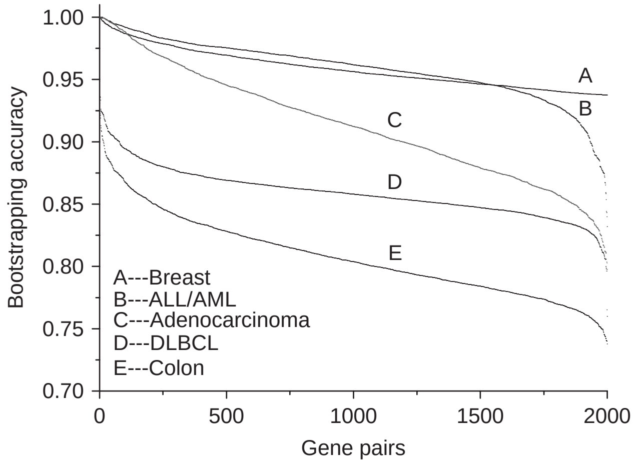

Fig. 4. The bootstrapping classification accuracy of the top-ranked 2000 classification models from breast, ALL/AML, adenocarcinoma, DLBCL and colon cancer data.

For colon cancer data set, the WSCA of all the classification models is lower than 100%. The classification model based on information gene pair, D14812 (human mRNA for ORF, complete cds) and H06524 (gelsolin precursor, plasma (human)), has the highest WSCA (93.55%), it has also the highest bootstrapping accuracy (91.59%). For DLBCL cancer data, there exist no classification model with 100% WSCA. The classification model based on information gene pair, GENE2760X_17204 and GENE3332X_13394, has the highest WSCA (97.62%). Classification model based on information gene pair, GENE1063X_16443 and GENE3330X_19288, has the highest bootstrapping accuracy (92.85%). Obviously, compared with the above three data sets, the colon and DLBCL cancer data sets have more complex structure.

Next, we used the method described in Sections 2.3 and 2.4 to find an optimal subset of classification models from the top-ranked classification models in group 1 or 2 and extracted feature subset from them (Table 1 lists the number of classification models in groups 1 and 2, $n_t = 2000$). When the length of chromosome is different, the performance of the extracted feature subset may be different. To extract excellent feature subset from classification models in groups 1 and 2, respectively, we performed a sequence of experiments in which we set the length of chromosome to be 10, 60, 110, 160 and 210, respectively. The experimental results on the five data sets are listed in Tables 2 and 3.Surprisingly, the best classification accuracy of the feature subsets extracted from the five data sets is all 100%. This indicates the proposed method can extract excellent features from microarray data. For colon cancer data set, when the length of chromosome is 10, the classification accuracy of feature subset extracted from group 1 is lower than 100%; maybe this attributes to the absentation of the feature subset with 100% classification accuracy among the top-ranked 10 classification models; when the length of chromosome is 160 and 210, respectively, the best WSCA of running 200 generation is lower

Table 1
The number of classification models in groups 1 and 2

Data	Group	Number	The description of the corresponding information gene pairs of classification models
Adeno-carcinoma	1	1700	Highly corrective in cancer samples
	2	300	Highly corrective in normal samples
ALL/AML	1	1486	Highly corrective in ALL samples
	2	514	Highly corrective in AML samples
Breast	1	4870	Highly corrective in BRCA1 mutation samples
	2	130	Highly corrective in BRCA2 and sporadic samples
Colon	1	1782	Highly corrective in normal samples
	2	218	Highly corrective in cancer samples
DLBCL	1	2000	Highly corrective in germinal center B-like DLBCL samples
	2	0	–

Table 2
The experimental results in group 1

Data	*LC*	OFV[a]	WSCA (%)	*Mg*/*MM*	*Fn*
Adeno-carcinoma	10	1.00608	100	0.602	4
	60	1.00847	100	0.840	18
	110	1.00810	100	0.802	23
	160	1.00750	100	0.743	42
	210	1.00736	100	0.728	51
ALL/AML	10	1.00424	100	0.418	4
	60	1.00828	100	0.819	7
	110	1.00843	100	0.835	17
	160	1.00824	100	0.816	32
	210	1.00831	100	0.823	45
Breast	10	1.00852	100	0.850	8
	60	1.00911	100	0.905	26
	110	1.00922	100	0.915	38
	160	1.00950	100	0.943	49
	210	1.00934	100	0.928	76
Colon	10	0.96781	96.774	–	3
	60	1.00254	100	0.245	6
	110	1.00166	100	0.159	28
	160	0.96782	96.774	–	30
	210	0.96782	96.774	–	43
DLBCL	10	1.00260	100	0.255	5
	60	1.00531	100	0.525	22
	110	1.00570	100	0.562	23
	160	1.00513	100	0.506	47
	210	1.00510	100	0.503	61

[a]Optimal fitness value.

Table 3
The experimental results in group 2

Data	*LC*	OFV[a]	WSCA (%)	*Mg*/*MM*	*Fn*
Adeno-carcinoma	10	1.00643	100	0.641	8
	60	1.00885	100	0.879	23
	110	1.00928	100	0.922	46
	160	1.00936	100	0.930	70
	210	1.00938	100	0.933	98
ALL/AML	10	1.00685	100	0.682	7
	60	1.00890	100	0.883	21
	110	1.00914	100	0.908	48
	160	1.00915	100	0.909	59
	210	1.00907	100	0.901	94
Breast	10	1.00830	100	0.826	6
	60	1.00974	100	0.970	33
	110	1.00971	100	0.965	43
	130	1.00975	100	0.970	59
Colon	10	1.00015	100	0.009	4
	60	1.00359	100	0.351	10
	110	1.00370	100	0.362	24
	160	1.00299	100	0.292	53
	210	1.00315	100	0.308	65

[a]Optimal fitness value.

than 100%; the reason for this result could be that the search space becomes larger and larger (for reaching the optimal feature subset, the program needs to run for a longer time) and the colon cancer data have more complex structure. From Tables 2 and 3, we can see that the *Mg*/*MM* of colon cancer data is the least. This indicates colon cancer data are more complex and difficult to classify.

3.3. Comparison with other methods

To further evaluate the performance of the proposed method, we compared the proposed method with the previously developed methods. Recent studies related to ALL/AM [19], adenocarcinoma [20], breast cancer [21–23] colon [7,9,24–26] and DLBCL [7,8] were chosen as targets for comparison. It is somewhat difficult to directly compare the proposed method with the above methods because they each employed different classifiers and evaluation strategies to test the features selected by them. Here we performed a comparison study exactly following their evaluation procedures.

Compared with colon and DLBCL data, adenocarcinoma, ALL/AML and breast cancer data have more simple structure. Many methods obtain 100% leave-one-out cross validation (LOOCV) accuracy in the three data sets. Wang et al. [19] identified 23 genes that can classify ALL/AML data 100%. Jaeger et al. [20] found 10 genes with 100% classification accuracy from adenocarcinoma cancer data. Cho et al. [22] obtained a subset of genes that produce zero misclassification on the breast cancer data. Lee et al.'s [23] method also produced zeros misclassification over three models that have 27, 17 and 10 genes, respectively. The proposed method and the above methods are not different in classification accuracy. However, in terms of efficiency and simplicity, one can argue that the proposed method is superior since the proposed classification model based on the best information gene pair only uses a pair of genes and an interpretable decision rule. The number of genes involved in the proposed classification model is significantly smaller than the number of genes in other methods. Although Cho et al. [21] also used a relatively small number of genes to classify breast cancer data set, the performance of their method is significantly worse than that of the proposed method. Table 4 lists the number of genes involved in different methods. Accurate diagnoses can be achieved using only a pair of genes. This makes cheap diagnostic microarray possible, because a microarray needs only two genes spotted on it instead of thousands of genes. This also reduces complexity of experiments and increases efficiency of disease diagnosis.

For colon cancer data, Xiong et al. [9] employed the sequential forward selection (SFS) algorithm and support vector machine (SVM) to selected feature genes from it; the best LOOCV accuracy of genes selected by them is 88.71%. Li et al. [25] combined SFS algorithm and Fisher's linear discriminant analysis to select feature genes that can discriminate between colon cancer data, and get 96.66% classification accuracy according the their evaluation method (training set included 95% of the

Table 4
The number of genes involved in different methods

Methods	The number of feature genes
(1) ALL/AML data set	
Wang et al. [19]	23
Proposed	2
(2) Adenocarcinoma	
Jaeger et al. [20]	10
Proposed	2
(3) Breast cancer data set	
Cho et al. [21]	3 (77.3%)[a]
Cho et al. [22]	21
Lee et al. [23]	10, 17, 27
Proposed	2

[a] The LOOCV accuracy of three feature genes selected by Cho et al. [21] is 77.3%.

total samples). Liu et al. [26] employed the ensemble neural network method to select feature genes and obtained 91.19% LOOCV accuracy. The above methods demonstrated the validity of their subset of genes by different evaluation methods after the completion of gene selection procedure. We followed their evaluation methods to further test the performance of our feature subset with theirs. Li et al. [7] combined GA and k-nearest neighbor (KNN) classifier to identify 50 most frequently selected genes that can discriminate between colon cancer data with higher classification accuracy. Li et al. [24] obtained 23 most frequently selected genes using an ensemble decision approach. Li et al.'s [7,24] methods are different from conventional wrapper methods. They first selected hundreds of gene subsets with higher classification ability using conventional wrapper methods, and then computed the frequency of each gene appearing in the selected hundreds of gene subsets. Finally, the genes with higher frequency are chosen as feature genes. Li et al.'s [7,24] tested the performance of their gene subsets using the KNN and SVM, respectively. Here we also compared the LOOCV results of their gene subsets with ours. Fig. 5 shows the comparison results.

For DLBCL data, Li et al. [7] identified 50 most frequently selected genes using GA and KNN classifier. Li et al. [8] also reported their average classification performance of the 7(5) genes obtained by combining GA and SVM (KNN) on the data set. Li et al. [7,8] evaluated the performance of their genes using different methods. In order to compare with their methods, we also employed the same classifiers and evaluation methods to test the performance our feature subset. Table 5 lists the comparison results.

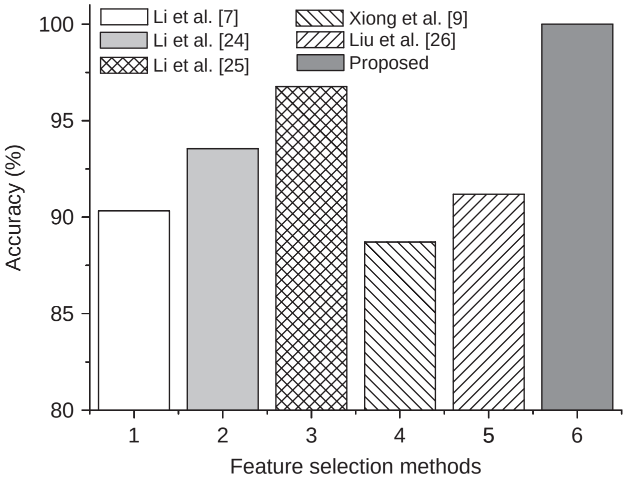

Fig. 5. The accuracy of features selected by six methods (colon cancer data).

Table 5
The classification accuracy (%) of the features selected by different methods on DLBCL data

Methods	SVM	KNN ($k = 1, 3, 5$)
Li et al. [7][a]	66.67	64.29/64.29/69.05
Li et al. [8][b]	92	95 ($k = 5$, GA–KNN)
Li et al. [8][b]	99 (GA–SVM)	65 ($k = 3$)
Proposed[c]	**100**	**100/100/100**

The best classification accuracy is highlighted in boldface.
[a] LOOCV accuracy.
[b] Five-fold cross validation accuracy.
[c] The LOOCV accuracy and five-fold cross validation accuracy.

From Fig. 5 and Table 5, we can see that the proposed method obtains satisfactory results on colon and DLBCL data sets. To our knowledge, this is the best result of the two data sets so far. Clearly, based on classification accuracy alone, our feature selection method is better, or at least comparable to other methods. Also, we find that the features selected by other methods can usually get better results on a certain classifier but may get worse results on the others. For example, Li et al. [8] extracted features by combining GA and SVM (GA–SVM). In order to compare their algorithm with the algorithm of combining GA and KNN (GA–KNN), they also combined GA and KNN to select feature genes. The features selected by GA–SVM have better classification performance on SVM classifier, while on KNN classifier they only obtain 65% classification accuracy. However, our features obtain best results on SVM and KNN. Therefore the features selected by the proposed method have better generalizability. It also indicates that the proposed method can effectively extract the intrinsic property of microarray data.

To compare the computational cost of different methods, we performed more experiments on colon cancer data with a complex structure. Feature gene selection methods generally fall into one of the two categories: filter and wrapper approaches. Filter methods usually first choose top k ($k = 50$ in this study) genes as feature genes and then employ classifiers, such as KNN [5] and SVM [5,27]). SNR and t-test are typical filter methods [3–5], they are chosen as targets for comparison. Wrapper methods usually combine classifiers and fast searching algorithm for finding an optimal set of genes [7–9]. Here we compared the computational costs of the proposed method with those of typical wrapper method: GA–KNN (GA–KNN indicates GA and KNN are combined for identifying an

Table 6
LOOCVl classification accuracy (%) and computational costs of different methods on colon cancer data

Methods	Time	*Acc* (%)
t-Test (KNN)[a]	2.02 s	91.94
SNR (KNN)[a]	2.06 s	91.94
GA–KNN	3.37 h	91.94
Li et al. [7]	33.86 h	93.54
Li et al. [8]	59.33 h	90.32
Proposed	25.54 h	100

[a] KNN classifier is used ($k = 5$).

optimal set of genes). Li et al.'s [7,24] methods are different from conventional wrapper methods. We also compared them with the proposed method. The LC of GA is 50 and other parameters of GA are the same as these in the above wrapper methods. All the algorithms are written using Matlab language and run on a PC with 2.4 GHz Pentium III CPU running windows XP. Table 6 lists the computational costs and LOOCV accuracy of different methods. It can be seen from Table 6 that filter methods occupy the least computational time and Li et al.'s[7,8] methods need most computational cost. Compared with GA–KNN, the proposed method needs more computational time and has better classification performance.

4. Conclusions

Information gene pairs are highly correlative in one type of sample; this indicates that they inter-regulate or get involved in the same biological processes such as cell cycle, metabolic pathway, signaling transduction pathway, and genetic regulatory pathway. The expression levels of one or/and two genes from a pair of information genes have significant changes in another type of sample means that the relations of the pair of information genes in different types of samples have changed. If the pair of information genes is still highly correlative in another type of sample, this indicates that inter-regulation intensity between the pair of genes has changed; if the pair of information genes is no longer highly correlative in another type of sample, this indicates that inter-regulation relation between the pair of genes has changed. Therefore, information gene pairs reflect the changes of the states of the cell, by which we might be more likely to capture invariant biological characteristics and extract more excellent features for accurate cancer classification. Based on the above idea, we construct the classification models based on information gene pairs and extract features from them. Experimental results on several microarray data sets have demonstrated that the proposed method performs well.

Acknowledgement

This work was supported in part by the National High Tech Development Project of China, the 863 Program (No.2007AA02Z302).

References

[1] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat. Med. 7 (2001) 673–679.

[2] P.J.S. Silva, R.F. Hashimoto, S. Kim, J. Barrera, L.O. Brandao, E. Suh, E.R. Dougherty, Feature selection algorithms to find strong genes, Pattern Recognition Lett. 26 (2005) 1444–1453.

[3] J.B. Welsh, P.P. Zarrinkar, L.M. Sapinoso, S.G. Kern, et al., Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer, Proc. Natl. Acad. Sci. USA 98 (2001) 1176–1181.

[4] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.

[5] J. Li, X. Tang, X. Li, A novel visualization classifier and its applications, Lecture Notes in Artificial Intelligence, vol. 3614, Springer, Berlin, 2005, pp. 1190–1199.

[6] X. Yan, M. Deng, W.K. Fung, M. Qian, Detecting differentially expressed genes by relative entropy, J. Theor. Biol. 234 (2005) 395–402.

[7] L. Li, T.A. Darden, C.R. Weinberg, A.J. Levine, L.G. Pedersen, Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method, Bioinformatics 17 (2001) 1131–1142.

[8] L. Li, W. Jiang, X. Li, K.L. Moser, Z. Guo, et al., A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset, Genomics 85 (2005) 16–23.

[9] M. Xiong, X. Fang, J. Zhao, Biomarker identification by feature wrappers, Genome Res. 11 (2001) 1878–1887.

[10] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511.

[11] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonu-cleotide arrays, Proc. Natl. Acad. Sci. USA 96 (1999) 6745–6750.

[12] D.A. Notterman, U. Alon, A.J. Sierk, A.J. Levine, Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays, Cancer Res. 61 (2001) 3124–3130.

[13] S. Theodoridis, K. Koutroumbas, Pattern Recognition, second ed., China Machine Press, 2003, pp. 387–390.

[14] J.H. Holland, Adaptation in Natural and Artificial Systems, The University of Michigan Press, Michigan, 1975.

[15] M. Srinivas, L.M. Patnaik, Genetic algorithms: a survey, IEEE Comput. 27 (1994) 17–26.

[16] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, et al., MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, Nat. Genet. 30 (2002) 41–47.

[17] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, et al., Gene expression profiles in hereditary breast cancer, N. Engl. J. Med. 344 (2001) 539–548.

[18] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, et al., Missing value estimation method for DNA microarray, Bioinformatics 17 (2001) 520–525.

[19] Y. Wang, F.S. Makedon, J.C. Ford, J. Pearlman, Hykgene: an hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data, Bioinformatics 21 (2005) 1530–1537.

[20] J. Jaeger, R. Sengupta, W.L. Ruzzo, Improved gene selection for classification of microarrays, Pac. Symp. Biocomput. (2003) 53–64.

[21] J.H. Cho, D. Lee, J.H. Park, I.B. Lee, New gene selection method for classification of cancer subtypes considering within-class variation, FEBS Lett. 551 (2003) 3–7.

[22] J.H. Cho, D. Lee, J.H. Park, I.B. Lee, Gene selection and classification from microarray data using kernel machine, FEBS Lett. 571 (2004) 93–98.

[23] K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucci, B.K. Mallick, Gene selection: a Bayesian variable selection approach, Bioinformatics 19 (2003) 90–97.

[24] X. Li, S. Rao, Y. Wang, B. Gong, Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling, Nucleic Acids Res. 32 (2004) 2685–2694.

[25] W. Li, M. Xiong, Tclass: tumor classification system based on gene expression profile, Bioinformatics 18 (2002) 325–326.

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

[27] L. Shen, E.C Tan, Kernel pls-SVM for microarray cancer classification, in: the Fourth Asia-Pacific Bioinformatics Conference, Taipei, Taiwan, 2006.