Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 27 (2006) 116–122
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Regulation probability method for gene selection

Hong-Qiang Wang [a,b,*], De-Shuang Huang [a]

[a] Intelligent Computation Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Science, P.O. Box 1130, Hefei, Anhui 230031, PR China
[b] Department of Automation, University of Science and Technology of China, Hefei 230027, PR China

Received 24 June 2004; received in revised form 5 July 2005
Available online 15 September 2005

Communicated by L. Goldfarb

Abstract

This paper proposes a novel method for gene selection. In the method, the gene regulation, an important mechanism of gene activities, is first introduced, and then the probabilities of gene regulation are estimated. These probabilities can be seen as the gene regulation information and can be used for gene selection. The applications to the leukemia dataset and the colon dataset suggest that our proposed method is effective, efficient, and competitive to the previous methods.

Keywords: DNA microarray dataset; Gene selection; Gene regulation probability; Probability estimation; Conditional probability

1. Introduction

Exploring the approaches for gene selection has become a very urgent and challenging task due to the difficulty and complication of dealing with the large scale of microarray data commonly containing 5000–10,000 gene measurements. A key goal of gene selection via different expression patterns from microarray data is to identify a handful of the responsible genes for certain events (say, certain diseases or certain types of tumors) so that it is easy and accurate for ones to design the suitable predictors for prognosing and diagnosing diseases. So far, there have been various machine leaning methods extensively applied to address the gene selection issue, e.g., clustering analysis (Eisen et al., 1998), signal-to-noise ratio (Golub et al., 1999), Fisher scores (Pavlidis et al., 2001), between-to-within variance (Dudoit et al., 2000) and various wrapper methods (Weston et al., 2000; Hastie et al., 2000; Pochet et al., 2004), etc. However, the outstanding feature of microarray data is the high degree of variability and noise pollution. The drawback of these methods is that they did not consider employing some suitable and necessary techniques to address these problems inherent in microarray data. Fortunately, some gene selection algorithms based on the probabilistic models have been proposed and the better results haven been obtained. For example, Baldic and Long (2001) have proposed a Bayesian approach to perform dimension reduction with a probit model; West et al. (2000) have presented a Bayesian framework for determining whether the observed differences in expression are significant or not. On the other hand, since a vast majority of variables remain hidden such that they must be inferred or integrated out by the probabilistic models, the usage of the statistical technique is necessary to the analysis of the microarray data (Baldic and Long, 2001; Lee et al., 2003).

This paper presents a novel algorithm for gene selection. In the algorithm, besides the information on the averages, the algorithm can extract the gene regulation information

* Corresponding author. Address: Department of Automation, University of Science and Technology of China, Hefei 230027, PR China. Fax: +86 551 5592751.
E-mail addresses: hqwang@iim.ac.cn (H.-Q. Wang), dshuang@iim.ac.cn (D.-S. Huang).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.07.007

from microarray datasets for gene selection. The usage of the regulation information in the proposed algorithm is similar to the previous methods for gene selection. Firstly, our proposed algorithm establishes the model of the gene regulation probability containing the gene regulation information, and then based on the model, the regulation probability is estimated by various statistical methods and is used to select genes. The successful applications to the leukemia and colon data show that our proposed algorithm is effective, efficient and comparable with the previous algorithms.

The remainder of this paper is organized as follows. In Section 2, we formulate the model and the computation of the gene regulation probability and give the detailed algorithm for gene selection. In Section 3, the proposed algorithm is applied to analyzing the leukemia and colon data. In particular, the comparisons of the performances with the previous algorithms are conducted in this section. Section 4 concludes this paper.

2. Main results

2.1. Modelling of gene regulation probabilities

Since the gene expression regulation under different biological situations can cause different phenotypes, e.g., some diseases, we can build the model of the probability of the gene expression regulation to identify the responsible genes for particular phenotypes, or diseases. Considering the binary case, we can assume that microarray data can be represented as a $G \times S$ expression matrix, A, with generic element, a_{gs}, representing the gene expression value of gene, g, in sample, s, i.e., the intensity value or transformed logarithmic value. Suppose that the two sample classes are denoted by the symbols, $Y = 1$ and -1, and contain M and N samples, respectively, and the target of each sample, s, can be denoted by $y_s \in \{1, -1\}$. Obviously, the equality, $M + N = S$, holds.

Biologically, the expression level of each gene in each sample should be in one of the below three states: the up-regulated state represented by $e = 1$, the down-regulated state by $e = -1$, and the non-significantly regulated state by $e = 0$. In what follows, the gene up-(down-)regulation states in each sample are statistically defined, respectively. Note that, the non-significantly regulated state can be defined as the state that does not belong to the up-regulated state or the down-regulated state.

Definition 1. Considering a gene, g, and a sample, s, we can define that the gene, g, is up-regulated in the sample, s, relative to the reference class of $Y = -1$ with N samples and the gene regulation state is called the up-regulation state, labelled by $e_{gs} = 1$, if

$$a_{gs} > a, \quad \forall a \in \{a_t, t = 1, 2, \ldots, \kappa\}, \tag{1}$$

where a_{gs} is the expression value of the gene, g, in the sample, s, the quantities $a_t \in \{a_{g(S-N+1)}, \ldots, a_{gS}\}$ and the parameter, κ, is an integer constraining to $0 < \kappa \leqslant N$.

Definition 2. Considering a gene, g, and a sample, s, we can define that the gene, g, is down-regulated in the sample, s, relative to the reference class of $Y = -1$ with N samples and the gene regulation state is called the down-regulation state, labelled by $e_{gs} = -1$, if

$$a_{gs} < a, \quad \forall a \in \{a_t, t = 1, 2, \ldots, \kappa\}, \tag{2}$$

where a_{gs} is the expression value of the gene, g, in the sample, s, the quantities $a_t \in \{a_{g(S-N+1)}, \ldots, a_{gS}\}$ and the parameter, κ, is an integer constraining to $0 < \kappa \leqslant N$.

We referred to the parameter, κ, as the gene regulation cutoff which can be determined by the number of samples in the reference class (here, i.e., the class of $Y = -1$) and has an effect on whether a gene can be justified to be regulated relative to the class or not. Setting $\kappa \leqslant N$ is to accommodate noise, variability or even errors inherent in the microarray data, because microarray data are very noisy and have high degree of variability such that the measured expression values vary irregularly (Newton et al., 2001). Commonly, we can preset $\kappa = \text{floor}(\eta N)$ where the cutoff coefficient $\eta \in (0.5, 1]$.

Based on Definitions 1 and 2, we can define such a regulation matrix that records all gene regulation states in each sample as follows:

Definition 3. The regulation matrix, B is such a matrix that has the same representation form as the microarray data matrix, but contains the generic element, $b_{gs} \in \{1, 0, -1\}$, recording the gene, g, regulation states in a sample, s, and determined by Definitions 1 and 2.

For the simplification of the computation, the regulation matrix, B, can be divided into two categories: the up-regulation matrix, B^+, and the down-regulation matrix, B^-. The up-regulation matrix, B^+, is used to record the gene up-regulation states in the samples of $Y = 1$ with respect to the class of $Y = -1$ and the gene down-regulation states in the samples of $Y = -1$ with respect to the class of $Y = 1$. The down-regulation matrix, B^-, is used to record the gene down-regulation states in the samples of $Y = 1$ with respect to the class of $Y = -1$ and the gene up-regulation states in the samples of $Y = -1$ with respect to the class of $Y = 1$. To obtain the regulation matrix, B^+ and B^-, firstly, the regulation states of each genes in each sample are computed in terms of Definitions 1 and 2 with respect to the corresponding class, and then each element of the regulation matrices, B^+ and B^-, is determined according to

For the up-regulation matrix B^+, each element, b^+_{gs}, is set by

$$b^+_{gs} = \begin{cases} 1 & e_{gs} = 1 \quad \text{and} \quad y_s = 1, \\ -1 & e_{gs} = -1 \quad \text{and} \quad y_s = -1, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

For the down-regulation matrix B^-, each element, b^-_{gs}, is set by

$$b_{gs}^{-}=\begin{cases}1 & e_{gs}=-1 \quad \text{and} \quad y_s=1,\\ -1 & e_{gs}=1 \quad \text{and} \quad y_s=-1,\\ 0 & \text{otherwise.}\end{cases} \tag{4}$$

In the statistical opinion, it is insufficient to infer the potential regulation rules of genes from only a single sample. For this reason, for each gene, we define two potential regulation events, i.e., the up-regulation event, denoted by $E=1$, and the down-regulation event, denoted by $E=-1$. The up-regulation event is the event that the gene is up-regulated in the class of $Y=1$ and is down-regulated in the class of $Y=-1$; while the down-regulation event is the event that the gene is down-regulated in the class of $Y=1$ and is up-regulated in the class of $Y=-1$. In order to identify the significant genes using the regulation information, our proposed algorithm uses the two probabilities of the above two regulation events, called the up-regulation probability and the down-regulation probability, respectively, to produce the criterion for the selection of genes. The two probabilities can be computed from the two regulation matrices, B^{+} and B^{-}, respectively. Because the absolutely independent regulation probabilities of genes cannot be got due to the limitation of the present human's knowledge on the gene regulation, we choose to use the conditional probabilities on a particular background situation, C, instead of the absolutely independent probabilities. In practice, the usage of the background condition, C, helps to filter the irrelevant genes to the class distinction in advance. So, we use the symbols, $P(E|C)$ ($P(E=1|C)$ and $P(E=-1|C)$ specially), to denote the above regulation probabilities of genes under a certain background condition, C. Our proposed algorithm utilizes the average information to produce the background condition, C. In what follows, the computation of the gene regulation probabilities, $P(E|C)$, using various statistical methods is formulated in detail.

2.2. The computation of gene regulation probabilities

Based on the above model of gene regulation probabilities, in this section we present the computation of the gene regulation probability using various statistical methods and finally summarize our proposed gene selection algorithm.

In Refs. (Golub et al., 1999; Baldic and Long, 2001), various probabilistic models based on the average, μ, and the deviation, σ, of gene expression levels were reported. However, in practice the average and the deviation contain highly correlated information. For simplicity and without the loss of the key information, our proposed algorithm produces the background condition, C, only by using the average quantity. The background condition, C, can filter out the genes that have relatively small differences in gene expression levels between two classes. After that, the left genes are collected as the preliminary gene set from which key genes will be selected using the information on the gene regulation. The background condition, C, can be represented by

$$C: |\mu_1-\mu_2| \geqslant \theta, \tag{5}$$

where μ_1 and μ_2 are the average values of the gene expression levels in the two sample classes, respectively, and θ is referred to as the expression difference cutoff. It is wise to refer to the minimum, $\theta_{\min}$, and the maximum, $\theta_{\max}$, of the average differences of gene expression levels between two classes to choose the value of the expression difference cutoff, θ, i.e., $\theta_{\min}<\theta<\theta_{\max}$. In what follows, the estimation of the regulation probabilities under the background condition, C, using the maximum likelihood and the Bayesian posterior estimation methods is formulated in detail.

Assuming that the regulation event, E, occurs by the probability, ψ, under the background context C, we have

$$P(E|C)=\psi, \tag{6}$$

where the unknown parameter, ψ, just is what we need to get. For each gene, if S regulation state values in the corresponding row of the gene regulation matrix, B, are looked at as S independent observations, then the occurring times, X, of the regulation event, E, will obey the binomial distribution assumption $b(S,\psi)$, i.e.,

$$P(X=x;\psi|C)=\binom{S}{x}\psi^{x}(1-\psi)^{S-x}, \quad x=0,1,2,\ldots,S, \tag{7}$$

where x can be got from the regulation matrix, $B(B^{+}$ or $B^{-})$, i.e.,

$$x=\begin{cases}\sum_{i=1}^{i=M} b_{gi}^{+}+\sum_{i=M+1}^{i=S}(-b_{gi}^{+}) & \text{for} \quad E=1,\\ \sum_{i=1}^{iM}(-b_{gi}^{-})+\sum_{i=M+1}^{i=S} b_{gi}^{-} & \text{for} \quad E=-1.\end{cases} \tag{8}$$

On the other hand, if, from Bayes assumption, the prior probability of ψ is supposed as $\pi(\psi)=U(0,1)$, Eq. (7) also can be thought as the joint probability of X and ψ. So, from Eq. (7), the marginal probability of X can be computed by

$$\begin{aligned} m(x|C) &= \int_0^1 P(X=x;\psi|C)\mathrm{d}\psi \\ &= \binom{S}{x}\frac{\Gamma(x+1)\Gamma(S-x+1)}{\Gamma(S+2)}. \end{aligned} \tag{9}$$

Finally, from Eqs. (7)–(9), our proposed algorithm obtains the identical estimation of the gene regulation probability by using the two different statistical methods, i.e., the Bayes posterior estimation and the maximum likelihood estimation methods. The estimated value of the gene regulation probability is

$$\hat{P}(E|C)=\frac{x}{S}. \tag{10}$$

So, in terms of Eq. (10), the two regulation probabilities of each gene can be obtained. Since only either of the two regulation events can occur to a single gene, the difference

between the two regulation probabilities will reflect the extent to which the gene is regulated. So, our proposed algorithm uses the absolute difference between the two probabilities to measure the importance of genes to the class distinction and as the criterion for selecting key genes.

Finally, our proposed gene selection algorithm can be summarized in detail as follows:

Begin

Step 1. Initialize the expression difference cutoff, θ, and the coefficient, η.

Step 2. Filter the genes by the background condition, $C: |\mu_1 - \mu_2| \geqslant \theta$.

Step 3. Generate the regulation matrix, B (B^+ and B^-), for the left genes in Step 2 by Eqs. (1)–(4).

Step 4. Compute the up-regulation and the down-regulation probability, i.e., $P(E=1|C)$ and $P(E=-1|C)$, for each of the left genes by $\hat{P}(E|C) = \hat{\psi} = \frac{x}{N}$.

Step 5. Rank and select genes in term of the absolute difference values between the two regulation probabilities, $P(E=1|C)$ and $P(E=-1|C)$, for each gene of the left genes.

End

3. Experimental results

In this section, we applied our proposed algorithm to the two publicly available microarray datasets: the leukemia and the colon. The leukemia dataset contains 72 tissue samples which are divided into two categories, i.e., the acute lymphoblastic leukemia (ALL) and the acute myeloid leukemia (AML), and each sample consists of expression values of 7129 gene variables. The colon dataset consists of 62 samples, of which 22 are normal and 40 tumor tissue, with the expression levels for 2000 genes. In our experiment, for assessing our proposed algorithm, each dataset was divided into two subsets: the training and test sets. For the leukemia data, the training set included 38 samples (27 ALL and 11 AML) and the test set included 34 samples (20 ALL and 14 AML). For the colon data, the training set included 40 samples (13 normal and 27 tumor) and the test set included 22 samples (9 normal and 13 tumor). These datasets are available at <http://www.genome.wi.mit.edu/MPR> and <http://www.molbio.princeton.edu/colon-data>, respectively.

3.1. Application to the leukemia data

To apply our proposed algorithm to the leukemia data, suppose that the subgroup of ALL is taken as the class of $Y=1$ and the subgroup of AML as the class of $Y=-1$. So, by the size of the training set, we have $M=27$ and $N=11$. The expression difference cutoff, θ, was set to one twentieth of the maximal absolute expression difference between the two subgroups, i.e., $\theta = 480.2343$. According to the results of many trials, the parameter, η, was chosen as 0.99. After our program was run on the training set, all the up-regulation probabilities, $P(E=1|C)$, and the down-regulation probabilities, $P(E=-1|C)$, for each of genes satisfying the condition, C, were carried out and the genes were ranked by the criterion of the absolute regulation probability difference. Fig. 1 displays the bar graphs of the distributions of the regulation probabilities with the interval of 0.1. Table 1 lists the 25 most significant genes as the ones with the largest regulation probability differences, where there are several genes, including the top two, that also belong to the set of 27 genes selected by Lee et al. (2003) and of 50 genes used by Golub et al. (1999) (these genes are reported with $*s$ and $\#s$ in Table 1, respectively).

To show the good performance of our proposed algorithm, the two well-known gene selection methods, i.e., the signal-to-noise ratio (SNR) method and the support vector machines recursive feature elimination (SVMRFE) method, were run on the same data and the obtained results were compared with ones by our proposed algorithm. Firstly, the highest correct classification rates on the test set were compared among the above three algorithms. For an extensive comparison, the two classifiers, i.e., support vector machines (SVM) and K-nearest neighborhood (KNN), were used to obtained the correct classification rates. Table 2 lists the classification accuracies and the number of the used genes for each of the above three gene selection methods. From Table 2, it can be found that our proposed algorithm can achieve the highest classification accuracies using

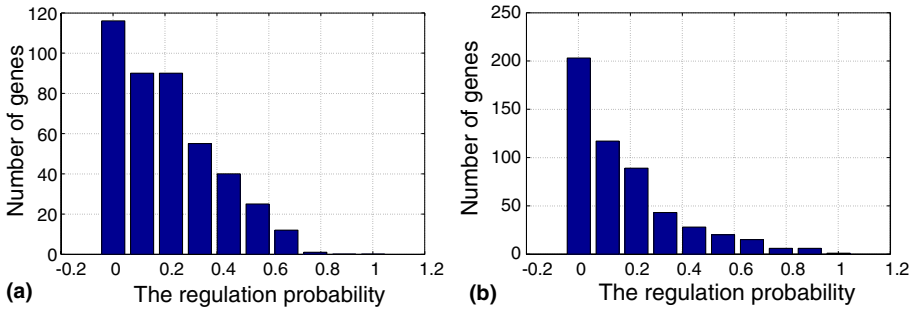

Fig. 1. The distributions of the regulation probabilities: (a) for the up-regulation probabilities; (b) for the down-regulation probabilities.

Table 1
Descriptions of the first 25 selected genes for the leukemia data

PD	Gene	
	Accession number	Description
1	X95735_at	Zyxin[*,#]
0.91	M27891_at	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)[*,#]
0.85	U50136_rna1_at	Leukotriene C4 synthase (LTC4S) gene
0.85	M27783_s_at	ELA2 Elastatse 2, neutrophil
0.81	D88422_at	CYSTATIN A[*]
0.81	M23197_at	CD33 CD33 antigen (differentiation antigen)[*,#]
0.81	Y12670_at	LEPR Leptin receptor
0.80	M16038_at	LYN V-yes-1 Yamaguchi sarcoma viral[*,#] related oncogene homolog
0.73	X66922_at	MYO-INOSITOL-1(OR 4)-MONOPHOSPHATASE
0.73	D26308_at	NADPH-flavin reductase
0.68	M19045_f_at	LYZ Lysozyme
0.68	D10495_at	PRKCD Protein kinase C, delta[*,#]
0.68	X14008_rna1_f_at	Lysozyme gene (EC 3.2.1.17)
0.67	U22376_cds2_s_at	C-myb gene extracted Human (c-myb) gene, complete primary cds,and five[*,#] complete alternatively spliced cds
0.67	U67963_at	Lysophospholipase homolog(HU-K5) mRNA
0.67	J03801_f_at	LYZ Lysozyme
0.66	M55150_at	FAH Fumarylacetoacetate
0.66	M63138_at	CTSD Cathepsin D (lysosomal aspartyl protease)
0.64	Z15115_at	TOP2B Topoisomerase (DNA) II beta (180kD)
0.63	M22960_at	PPGB Protective protein for beta-galactosidase (galactosialidosis)[*]
0.62	X85116_rna1_s_at	Epb72 gene exon 1
0.61	M11507_3_at	AFFX-HUMTFRR (endogenous control)
0.61	U82759_at	GB DEF = Homeodomain protein HoxA9 mRNA
0.61	X51521_at	VIL2 Villin 2 (ezrin)
0.60	HG1612-HT1612_at	Macmarcks[*]

PD: Probability difference.
[*] Also belonging to the set of 27 genes selected by Lee et al. (2003).
[#] Also belonging to the set of 50 genes used by Golub et al. (1999).

Table 2
Comparison of the classification performance with the previous methods for the leukemia dataset

	KNN classifier		SVM classifier	
	Accuracy (%)	No. of genes	Accuracy (%)	No. of genes
Our algorithm	97.09	8	100	29
SNR	97.09	7	97.09	18
SVMRFE	94.12	15	100	8

quite a few genes, regardless of the used classifiers. Secondly, the change curves of classification accuracies with different numbers of the used genes were compared. For this comparison, the classification accuracies were obtained by the KNN classifier. The resulting changing curves for each gene selection method are shown in Fig. 2. From Fig. 2, it can be found that our proposed algorithm is better than others both in stability and in accuracy. In addition, the comparison of the classification performance of a single gene was conducted based on the KNN classifier. Table 3 reports the classification accuracies of each single gene of the first 25 genes ranked by the three gene selection method. From Table 3, it can be found that the accuracies for our algorithm are better than those for the SVMRFE method and competitive with those for the SNR method.

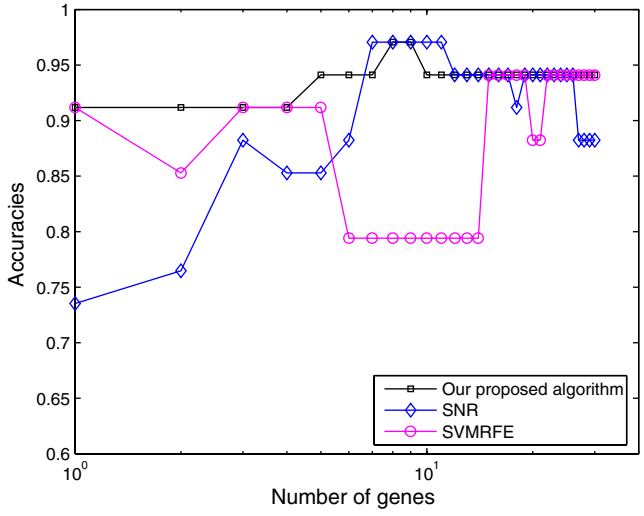

Fig. 2. Comparison of classification accuracies with the SNR and SVMRFE methods.

By the above multiple comparisons, we can draw a conclusion that our proposed algorithm is competitive with the SNR and SVMRFE methods for the leukemia data.

In our algorithm, the regulation cutoff coefficient, η, is designed to accommodate noise and avoid the harm of high

Table 3
Comparison of classification accuracies by a single gene with the SNR and SVMRFE methods for the leukemia data

Our proposed algorithm	SNR	SVM-RFE
0.91176	0.73529	0.82353
0.94118	0.79412	0.52941
0.79412	0.91176	0.44118
0.64706	0.67647	0.76471
0.85294	0.79412	0.82353
1	1	0.94118
0.61765	0.73529	0.64706
0.79412	0.61765	0.73529
0.70588	0.70588	0.52941
0.64706	0.73529	0.38235
0.61765	0.76471	0.55882
0.73529	0.94118	0.82353
0.79412	0.82353	0.41176
0.67647	0.79412	0.58824
0.58824	0.52941	0.58824
0.73529	0.55882	0.55882
0.73529	0.82353	0.61765
0.82353	0.76471	0.55882
0.85294	0.64706	0.79412
0.76471	0.61765	0.58824

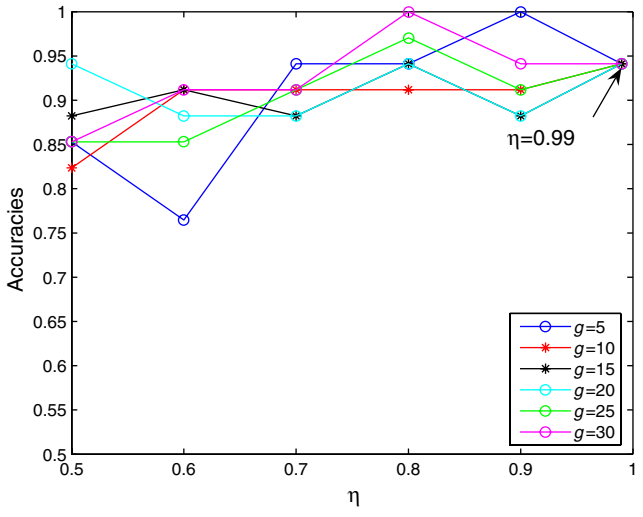

Fig. 3. Change curves of classification accuracies with the cutoff coefficient η.

variability inherent in microarray data. It is necessary for us to test the influence of the parameter on the performance of the proposed algorithm. To this end, our algorithm was repeatedly run on the leukemia data under different values of η. The curves of the relationship between classification accuracies and η were obtained and were shown in Fig. 3. For each curve in Fig. 3, the number, g, of the used genes was fixed and the classification accuracies were obtained through the KNN classifier classifying the test set. From Fig. 3, it can be seen that, all the curves indicate that the classification accuracies can reach the peaks when $\eta \in [0.8, 1)$. On the other hand, the stability appears to increase with η. In particular, when $\eta = 0.99$, all the curves join, that is to say, the classification accuracies change with g no longer.

3.2. Application to the colon data

The colon dataset is also popular by researchers. For this dataset, Alon et al. (1999) clustered 62 samples into two clusters: one consisting of 35 tumor and 3 normal and another consisting of 19 normal and 5 tumor tissues. Here, this dataset was analyzed by our proposed algorithm. In experiments, normal tissue samples were taken as the class of $Y = 1$ and tumor tissue samples as the class of $Y = -1$. So, the parameters, M and N, were equal to 13 and 27, respectively. Using the similar techniques in the first application, other parameters can be chosen as: $\theta = 278.98$ and $\eta = 0.99$. As a result, the genes were ranked by our proposed algorithm. Table 4 lists the 10 genes that have the largest regulation probability differences. Note that, because the "M76378" gene was repeatedly hybridized in the hybridization experiment, our proposed algorithm repeatedly ranked the gene and placed it at the continuous locations: the fourth, fiveth and sixth rows in Table 4. This result is accordant with the fact, confirming that our proposed algorithm is effective and can obtain the reasonable regulation probabilities. Next, the selected genes were used to classify the test set and the classification accuracies were compared with ones for the SNR and SVMRFE methods. Table 5 reports the classification accuracies and the number of the used genes. From Table 5, it can be seen that our proposed algorithm improves the classification accuracies and the number of the used genes are reduced, regardless of the used classifiers. In addition, Table 6 compares the classification accuracies of one single

Table 4
Descriptions of the first 10 selected genes for the colon data

PD[a]	Gene	
	Accession number	Description
0.525	R87126	Myosin heavy chain, nonmuscle (*Gallus gallus*)
0.525	H43887	Complement factor D precursor (*Homo sapiens*)
0.4	R28373	Hemoglobin beta chain (human)
0.4	M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6
0.375	M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6
0.35	M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6
0.35	T92451	Tropomyosin, fibroblast and epithelial Muscle-type (human)
0.35	J02854	Myosin regulatory light chain 2, smooth muscle Isoform (human); contains element TAR1 repetitive element
0.3	M63391	Human desmin gene, complete cds
0.3	T59162	74635 selenium-binding protein (Mus musculus)

[a] PD: Probability difference.

Table 5
Comparison of the classification performance with the previous methods for the colon dataset

	KNN classifier		SVM classifier	
	Accuracy (%)	No. of genes	Accuracy (%)	No. of genes
Our algorithm	81.82	6	81.82	8
SNR	77.27	1	77.27	5
SVMRFE	81.82	14	81.82	14

Table 6
Comparison of classification accuracies by a single gene with the SNR and SVMRFE methods for the colon data

Our proposed algorithm	SNR	SVM-RFE
0.72727	0.77273	0.72727
0.68182	0.59091	0.68182
0.5	0.72727	0.45455
0.72727	0.68182	0.54545
0.68182	0.59091	0.59091
0.72727	0.72727	0.54545
0.68182	0.5	0.59091
0.77273	0.72727	0.54545
0.77273	0.45455	0.63636
0.40909	0.59091	0.5

gene for the first 10 genes among the three gene selection methods. From Table 6, it can be found that our algorithm appears to outperform both the SVMRFE method and the SNR method.

4. Conclusions

This paper proposed a novel gene selection algorithm based on gene regulation probabilities. After the model of the gene regulation probability is established, our algorithm employs several statistical methods to estimate the probabilities of the up-regulation and down-regulation events for each gene and computes the absolute differences between the two probabilities for ranking and selecting the significant genes. Besides the information on the averages, the algorithm can extract the gene regulation information from microarray datasets for gene selection. The usage of the gene regulation information in the proposed algorithm is similar to the previous methods for gene selection. In the discussed applications to the leukemia and the colon data, it is shown that our proposed algorithm can identify key genes. Based on the *K*-nearest neighborhood and SVM classifiers, our proposed algorithm is comparable in the classification performance with the SNR and SVMRFE methods. The comparison results show that the proposed algorithm is stable and the genes are found with quite high classification accuracies. In addition, the role of the parameter, η, in our algorithm is discussed and it is suggested that it should be chosen between 0.80 and 0.99. Future research work will include extensive research on the conditional probability of the gene regulation events and more applications to various microarray datasets. In addition, the computation of probabilities of a single-gene regulation will be extended into the case of the multi-gene regulation for better selection of key genes.

References

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

Baldic, P., Long, A.D., 2001. A Bayesian framework for the analysis of microarray expression data: Regularized *t*-test and statistical inferences of gene changes. Bioinformatics 17, 509–519.

Dudoit, S., Fridlyand, J., Speed, T., 2000. Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report 576. Available from: <http://www.stat.berkeley.edu/sandrine/tecrep/576.pdf>.

Eisen, M., Spellman, P., Brown, P., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95, 14863–14868.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gassenbeek, M., Mesirov, J.P., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E., 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–538.

Hastie, T., Tibshirani, R., Eisen, M.B., et al., 2000. Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol. 1 (2), research0003.1–0003.21.

Lee, K.E., Sha, N.J., Dougherty, E.R., Vannucci, M., Mallick, B.K., 2003. Gene selection: A Bayesian variable selection approach. Bio-informatics 19, 90–97.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., Tsui, K.W., 2001. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. Comput. Biol. 8 (1), 37–52.

Pavlidis, P., Weston, J., Cai, J., Grundy, W.N., 2001. Gene functional classification from heterogeneous data. In: Proc. 5th Internat. Conf. on Comput. Mol. Biol.

Pochet, N., Smet, F.D., Suykens, J.A., Moor, B.D., 2004. Systematic benchmarking of microarray data classification: Assessing the role of non-linearity and dimensionality reduction. Bioinformatics 20, 3185–3195.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., Vapnik, V., 2000. Feature selection for SVMs. Adv. Neural Inform. Process. Syst., 13.

West, M., Nevins, J.R., Marks, J.R., Spang, R., Zuzan, H., 2000. Bayesian regression analysis in the large *p*, small *n* paradigm with application in DNA microarray studies. Technical Report. Duke University.