New Variable Selection Method Using Interval Segmentation Purity with Application to Blockwise Kernel Transform Support Vector Machine Classification of High-Dimensional Microarray Data

Li-Juan Tang, Wen Du, Hai-Yan Fu, Jian-Hui Jiang,* Hai-Long Wu, Guo-Li Shen, and Ru-Qin Yu*

State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, P. R. China

Received January 24, 2009

One problem with discriminant analysis of microarray data is representation of each sample by a large number of genes that are possibly irrelevant, insignificant, or redundant. Methods of variable selection are, therefore, of great significance in microarray data analysis. A new method for key gene selection has been proposed on the basis of interval segmentation purity that is defined as the purity of samples belonging to a certain class in intervals segmented by a mode search algorithm. This method identifies key variables most discriminative for each class, which offers possibility of unraveling the biological implication of selected genes. A salient advantage of the new strategy over existing methods is the capability of selecting genes that, though possibly exhibit a multimodal distribution, are the most discriminative for the classes of interest, considering that the expression levels of some genes may reflect systematic difference in within-class samples derived from different pathogenic mechanisms. On the basis of the key genes selected for individual classes, a support vector machine with block-wise kernel transform is developed for the classification of different classes. The combination of the proposed gene mining approach with support vector machine is demonstrated in cancer classification using two public data sets. The results reveal that significant genes have been identified for each class, and the classification model shows satisfactory performance in training and prediction for both data sets.

1. INTRODUCTION

Microarrays allow scientists to analyze expression of many genes in a single experiment in very quick and efficient manner. Such technologies are widely used to investigate the fundamental aspects of growth and development, as well as to explore the underlying genetic causes of many human diseases. Usually, thousands of genes are simultaneously detected on one microarray chip, followed by bioinformatic data analysis, such as classification or regression. One problem with discriminant analysis of microarray data is representation of each sample by a large number of genes that are possibly irrelevant, insignificant or redundant. Methods of variable selection are, therefore, of great significance in microarray data analysis, which is helpful for thorough understanding of the pathogenesis and accurate diagnosis and therapy of the diseases.

At present, a great deal of variable selection methods have been applied to identify differentially expressed genes. Many methods, such as *t*-statistic,[1] the ratio of between classes sum of squares to within class sum of squares (BSS/WSS),[2] nearest shrunken centroid,[3] and tree-based emerging patterns discovery[4] select key genes via the ascertainment of expression average and deviation of the genes between classes based on a presumption that the expression of the genes within one class is unimodal. Nevertheless, it is common that a single disease state may have multiple subtypes derived from varying pathogenic mechanisms that require different treatments. This implies that genes for samples within a single class may reflect systematically differential expression levels. In other words, expression pattern of a gene for a class may show a multimodal distribution with intrinsic clusters.[5,6] In such cases, the average expression for the class may show small discrepancy from those for the other classes relative to the within-class variations, but the expression patterns for the class can form clusters, or intervals, that all significantly separate from those for the other classes. As an example, Figure 1 illustrates the expression patterns of two genes for two classes. The expression pattern of gene 1 for Class 1 shows a multimodal distribution with intrinsic clusters. As most methods treating all the samples in class 1 as a whole, gene 2 would be considered more discriminative than gene 1 in distinguishing the two simulated classes. For example, the absolute values of the *t*-statistic of the two classes are 0.2203 for gene 1 and 12.2943 for gene 2, denoting that on gene 2 the two classes are more different from each other than on gene 1, whereas it is obvious that on gene 1 the two classes have more clear distribution intervals than on gene 2 if treating Class 1 as two clusters. Consequently, this kind of gene, for example, gene 1, is very discriminative for the classes of interest. However, these genes would be ignored by conventional variable selection methods. Thus, in microarray data analysis, it seems neces-

* To whom correspondence should be addressed. E-mail: rqyu@hnu.cn, jianhuijiang@hnu.cn. Tel: +86-731-8822577. Fax: +86-731-8822782.

10.1021/ci900032q CCC: $40.75
Published on Web 07/31/2009

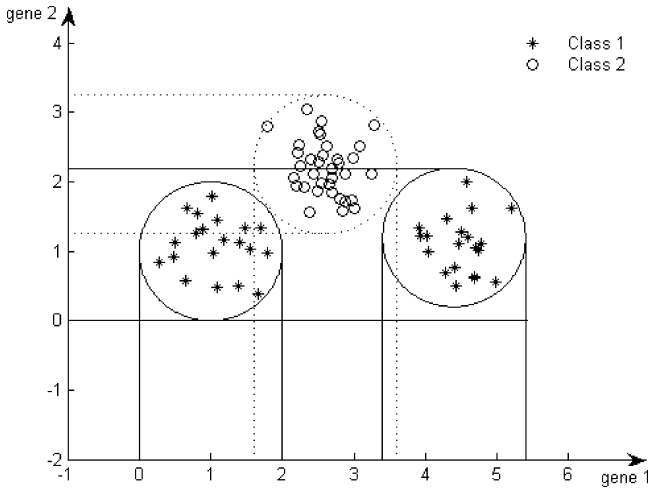

Figure 1. Simulated expression patterns of two genes for two classes while the expression pattern of one gene for one class shows a multimodal distribution with intrinsic clusters.

sary and plausible to develop strategies that can effectively locate such key genes.

To identify the key genes probably exhibiting multimodal expression patterns in one class, a new variable selection method has been proposed based on interval segmentation purity (ISP) as defined in the following parts. Intuitively, a gene discriminative for a class of interest implies that the probability density function of its expression level for the class is significantly different from those of the others, that is, its expression levels for the class are centralized at one or several intervals where rarely appear the expression levels for the other classes, as illustrated in Figure 1. In other words, there are some intervals where the samples from the class of interest are "pure" with respect to other classes, and these intervals can be regarded to have adequate purity of samples belonging to the class of interest. In this sense, the discriminative power of a gene for one class can be measured by the "purity" defined by the aforementioned reasoning for the intervals where the expression levels of the gene for the class are centralized. On the basis of the concept, mean shift algrothm[7] is invoked to locate the intervals where the expression levels of a gene for one class are centralized. Mean shift[7−9] is a nonparametric clustering technique that does not require any prior knowledge of the number and the shape of clusters. When used for clustering the data represented by a single variable, mean shift is equivalent to segmenting the variable axis into several intervals via a model search process, and the algorithm can be implemented in a parallel, computationally efficient manner, affording a highly desirable technique in handling multidimensional microarray data. With intervals segmented via such a mode search strategy, the ISP as quantitatively defined in the present paper (vide infra) can be estimated for each gene, and the key genes can be identified for each class of interest. A salient feature of this method is to identify key variables most discriminative for each class. This offers possibility of unraveling the biological implication of selected genes. Also, the new strategy furnishes the advantage over existing methods in the capability of selecting discriminative genes, whether the genes are up-regulated, down-regulated, or multimodal. Moreover, to exploit the information contained in the genes discriminative for each specific class, support vector machine (SVM)[10,11] with block-wise kernel transform (BKT) has been employed. Kernel transform is frequently employed by SVM to solve a nonlinear problem. Du and co-workers have demonstrated that it is an effective strategy to allocate variables into several blocks using a certain criterion followed by extracting the features of each block separately.[12] Unlike a conventional kernel transform that is performed for all selected variables, BKT is applied to variable subsets that are each selected for a specific class of interest. This allows selective extraction of discriminant information from the genes selected for the class of interest, offering the possibility of eliminating the interference from redundant genes from the other classes. Thus, BKT gives an ideal feature mapping for the selected genes, which results in feature variables with larger values for the class of interest and smaller values for the others. The combination of the proposed gene mining approach with support vector machine is evaluated in cancer classification using two public data sets, the small, round blue-cell tumors (SRBCTs) of childhood data[13] and a complex combination of clinical and histopathological data (GCM).[14]

2. THEORY

2.1. Interval Segmentation Purity (ISP)-Based Variable Selection Algorithm. The idea of the proposed variable selection method is to segment each variable axis into intervals or clusters where the expression level values are centralized, followed by the evaluation of the purity of these clusters for a class of interest. To disclose the natural clusters of the data, a straightforward approach is a mode search algorithm, by which the clusters are defined as the unimodal distribution of the probability density function.[15] Mean shift[7−9] is a computationally efficient algorithm for mode search based clustering. This procedure is essentially a gradient ascent algorithm started from every data point. Then, all data points located in the same unimodal intervals, that is, in the same cluster, converge at the same mode within a specified error level δ (a parameter to be set in the algorithm, in the present study $\delta = 10^{-4}$), so the cluster membership of a data point can be defined by the mode at which it converges. Suppose that a microarray data set has P genes and N samples with the entries x_{pn} ($p = 1, ..., P$; $n = 1, ..., N$) representing the expression level of the nth sample, $\mathbf{x}_n$, on the pth gene. And let K be a flat kernel that is the characteristic function with a certain analysis bandwidth h,

$$K(x) = \begin{cases} 1 & \text{if} ||x|| \leq h \\ 0 & \text{if} ||x|| > h \end{cases} \tag{1}$$

The mean shift vector in a simple one-dimensional case (on one gene, that is, on gene p) can be expressed as

$$m(x) = \frac{\sum_{n=1}^{N} x_{pn} K(x - x_{pn})}{\sum_{n=1}^{N} K(x - x_{pn})} - x \tag{2}$$

where x is a data point in a one-dimensional data space, starting from an arbitrary data point, x_{pn}, for $n = 1, ..., N$; the analysis bandwidth h is a positive value which can be determined by sensitivity analysis.[16,17] When the function m is applied to the original point, x, it results in a new

position, x^s; this process can be repeated until $|m(x)| < \delta$ and an iterative procedure is defined in this way:

$$x^{s+1} = x^s + m(x^s) \tag{3}$$

For a kernel with a monotonically decreasing profile, convergence of x^s ($s = 1, ..., S$) can be proven. The iterative mean shift procedure is, in essence, a gradient ascent method where the step size is initially large and decreases toward convergence. A prominent merit of this algorithm is that it does not require any prior knowledge of the number and the shape of clusters, approximating the true distribution of the data. When used for clustering the data represented by a single variable, the mean shift algorithm results in segmentation of the variable axis into several intervals (where the modes are centralized).

Suppose that the mean shift algorithm segments the pth variable (gene) axis into Q intervals (for distinct variables, p, the value of Q might change for different distribution patterns associated with samples on different genes), and there are n_q ($q = 1, ..., Q$) samples from different classes, including $n_q(k)$ samples from class k, converging at the same mode located in the qth interval, the ISP score of gene p for class k ($k = 1, ..., K$), ISP(k), is defined as follows:

$$\mathrm{ISP}(k) = \sum_{q=1}^{Q} w_q \frac{n_q(k)}{n_q} = \sum_{q=1}^{Q} \frac{n_q(k)}{n(k)} \frac{n_q(k)}{n_q} (k = 1, ..., K) \tag{4}$$

where w_q denotes the weighting factor determined by the purity of the qth interval to the total ISP(k), which is determined by the percent of sample number belonging to class k in the qth interval to total sample number of class k, that is, $w_q = n_q(k)/n(k)$, where $n(k)$ symbolizes the number of samples in class k, that is, $\sum_{q=1}^{Q} n_q(k) = n(k)$ and $\sum_{k=1}^{K} n(k) = N$. Based on the definition, large values of ISP(k) imply that expression levels of gene p comprise intervals in which most of samples come purely from class k, indicating that this gene is useful for discriminating class k from the others. Thus, the genes with the largest $ISP(k)$ can be identified as the key genes for class k. In general cases, several key genes are selected for each class, and the optimal number, say J, of key genes can be determined by trial-and-error procedure or cross validation with varying number of key genes. In the present study, the same optimal number of genes, J, has been used for each class. Also, for different classes, J would be different. But, further experiments have demonstrated that it improved the classification results little. The simplest way is to choose the same least number of key genes when a model reaches desirable classification accuracy for the training set. A set of L genes ($L = J \times K$) will be used for training a SVM model.

Note that the mean shift algorithm not only gives the interval segmentation of each gene for the evaluation of ISP but also allows direct determination of the cluster centers for each class using the mode thus located. This offers an additional benefit for the following SVM with BKT.

2.2. Support Vector Machine with Block-wise Kernel Transform (BKT-SVM). SVM[10,11] is a promising machine learning technique with comprehensive theoretical foundation. To solve the problem of nonlinearity, generally SVM uses a device called kernel mapping to transform the data from the original variables to a feature space, in which the model becomes linear, followed by a linear SVM technique to get the solution. Gaussian radial basis function transform is frequently utilized if the knowledge of a problem dealt with is lacking. When implementing the kernel mapping, conventional SVM presents a global transform manner, applying kernel transform to all the input variables, for nonlinear feature extraction. Because the proposed gene mining approach selects a series of gene sets that have respectively discriminative information for each individual class, using a global kernel transform as in standard SVM would not be a good strategy here. Therefore, a Gaussian BKT is proposed and used for SVM in the present study to transform the data from the original variables space to the feature one in which the feature of each class could be more effectively extracted.

For a classification problem with K classes, there would be K subsets of variables identified by ISP-based method. In the Gaussian BKT, each subset of variables forms a block, resulting in K blocks each including J variables (genes). Then, a sample in the kth block can be represented as $\mathbf{x}_n(k)$ ($n = 1, ..., N$; $k = 1, ..., K$), denoting the J-dimensional expression profile vector for the nth sample. Giving $\mathbf{X}(k)$ as the input matrix (the expression profiles of N samples) in the kth block, the output of D kernel centers in the kth block, $\mathbf{O}(k)$, can be represented as follows:

$$\mathbf{O}(k) = \exp\left(-\frac{||\mathbf{X}(k) - \mathbf{C}(k)||^2}{2\boldsymbol{\sigma}(k)^2}\right) \quad (k = 1, ..., K) \tag{5}$$

where $\boldsymbol{\sigma}(k) = (\sigma_1(k), .., \sigma_D(k))$ and $\mathbf{C}(k) = (\mathbf{c}_1(k), ..., \mathbf{c}_D(k))$ are the kernel widths and centers, respectively, of the kth block. The transform centers $\mathbf{C}(k)$ in the kth block are the centers of the clusters for samples belonging to class k, which is immediately determined by combining the modes where the entries of $\mathbf{x}_n(k)$ ($n = 1, ..., n(k)$), all training samples in class k, converge in mean shift clustering. In addition, the number of kernel center D is automatically determined, which might vary for different blocks. For an example, if there are ten samples from class k described by two identified variables in the training set, converging at one mode $c_{1,1}$ on the first variable and two modes $c_{2,1}$ and $c_{2,2}$ on the second variable, then there will be two clusters of the ten samples with two cluster centers located at $\mathbf{c}_1(k) = (c_{1,1}, c_{2,1})$ and $\mathbf{c}_2(k) = (c_{1,1}, c_{2,2})$. The output $\mathbf{O} = (\mathbf{O}(1), ..., \mathbf{O}(K))$ can be regarded as a set combined by feature variables extracted from all blocks of variables using a Gaussian BKT. The kernel width $\sigma_d(k)$ is calculated by the distance between $\mathbf{c}_d(k)$ and the mean vector of the class that is the nearest to $\mathbf{c}_d(k)$ among all classes other than class k.[18] Note that samples belonging to class k have smaller distance from one of the kernel centers of class k than from centers of other classes, then the BKT for class k is expected to give larger output values for samples belonging to class k than those from other classes. As the output values afford more immediate information for discriminating the classes than the selected genes, it is clear that the proposed BKT is very effective in feature extraction and is able to substantially enhance the classification performance of SVM. In SVM learning, one also needs to encode a K-dimensional vector of category label for the nth sample, $\mathbf{y}_n = (y_{1,n}, ..., y_{K,n})$ with 1 in the kth coordinate and 0 elsewhere if the sample falls into class k ($k = 1, ..., K$). A

regression-based SVM classifier is employed to model the following relationship between the extracted feature variables **O** and the category label, $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_K)$ of the samples:

$$\mathbf{Y} = \mathbf{W} \times \mathbf{O} + \mathbf{e} \quad (6)$$

where $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_K)$ and $\mathbf{e} = (e_1, ..., e_K)$ are the weight matrix and the thresholds, respectively, for the model of class k ($k = 1, ..., K$). For a regression of $\mathbf{y}_k$

$$\mathbf{y}_k = \mathbf{w}_k^T \times \mathbf{O} + e_k \, (k = 1, ..., K) \quad (7)$$

The optimal regression functions are given by the minimizing

$$\frac{1}{2}\mathbf{w}_K^{\mathrm{T}}\mathbf{w}_k + C\frac{1}{N}\sum_{n=1}^{N} L_\varepsilon(y_{kn} - \hat{y}_{kn}) \quad (8)$$

where

$$L_\varepsilon(y_{kn} - \hat{y}_{kn}) = \begin{cases} |\hat{y}_{kn} - y_{kn}| - \varepsilon & |\hat{y}_{kn} - y_{kn}| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

is the ε-insensitive loss function measuring the error between the given observations $\mathbf{y}_k$ and the estimated ones $\hat{\mathbf{y}}_k$ ($k = 1, ..., K$) and ε is the tolerance zone. The vector $\mathbf{w}_k$ is a normal vector, and $1/2(\mathbf{w}_k^{\mathrm{T}}\mathbf{w}_k)$ is used as a measurement of the model complexity, defining the structure risk of a SVM model. A penalty constant C is introduced to determine the trade-off between the empirical error and the model complexity. The optimal model parameters are estimated by solving a quadratic programming problem, in which the penalty constant C and the tolerance zone ε are determined by sensitivity analysis. With estimated model parameters, the classifier can be constructed immediately according to the estimated values $\hat{\mathbf{y}}_k (k = 1, ..., K)$ for any training or prediction samples: the sample is allocated into the hth class, where h is the number at which $\hat{y}_{hn}$ reaches the maximum among all $\hat{y}_{kn}$ values ($k = 1, .., K$).

3. RESULTS AND DISCUSSION

3.1. Small, Round Blue-Cell Tumor Data. In this paper, the data set of the small, round blue-cell tumors (SRBCTs) of childhood[13] is revised as a four-class problem with 83 samples, including neuroblastoma, rhabdomyosarcoma, Burkitt lymphoma and the Ewing family of tumors. Discrimination of these four types has presented a challenge because of their similar appearances in routine histology. In 2001, Khan and co-workers successfully monitored the gene expression profiles of 6567 genes for these four types of malignancies using cDNA microarrays and reduced the number of genes to 2308 by quality filtering for a minimal level of expression.[13]

First, the stability of the ISP-based variable selection method was tested through leave-one-out cross validation (CV). Each time one sample was removed from the data set and the remaining made up the reduced data set. The percentages of the selected top genes ranked by their ISP scores that are common either using the original data set or the reduced data sets were recorded. Total 83 reduced data sets were tried and the average percentage was taken to measure the stability for ISP method. In each computation, the same number of top ranked genes is detected for different classes. Table 1 collects the stability results of ISP when 4, 10, 30, 50, or 100 genes were detected in each class. From these results, it is clear that ISP method shows desirable stability. The top five genes with the highest reselection percentages in each class are listed in Table 2. These genes were reported to encode functional proteins responsible for transcription, development, metabolism and structure associated with the SRBCT disease.[19−23] Figure 2 depicts the expression levels of the highest reselection percentages for SRBCT. It is clear from the heat map that the 20 genes selected by ISP method are very informative in discriminating the four classes.

Then, the ISP method was examined in handling the SRBCT data in comparison with other two commonly used ones, BSS/WSS[2] and Wilcoxon rank sum test (WRST).[24] The genes selected by these methods were used to establish the model using different classification techniques, standard kernel transform-SVM, and the proposed BKT-SVM. To avoid a fortuitous choice of the training and the test sets, the original data set was randomly divided into five completely independent parts of roughly equal size. In each

Table 1. Stability Results of ISP Method on Identifying Key Genes for Each Class in SRBCT Data When Different Optimal Numbers of Genes Are Detected

no. of optimal genes selected	EWS[a]	BL[a]	NB[a]	rms[a]
4	0.9699	0.8494	0.9729	0.8584
10	0.9554	0.8277	0.9217	0.9145
30	0.9032	0.9201	0.8847	0.8928
50	0.8745	0.9224	0.8402	0.9176
100	0.8773	0.9146	0.8621	0.8871

[a] EWS, Ewing family of tumors; BL, Burkitt lymphoma; NB, neuroblastoma; rms, rhabdomyosarcoma.

Table 2. Top Five Genes with the Highest Reselection Percentages for Each Class in SRBCT Data, Ranked According to the LOOCV Result Obtained by ISP Gene Selection Method

class	image id[a]	gene description
EWS[b]	357031	tumor necrosis factor, alpha-induced protein 6
	1435862	antigen identified by monoclonal antibodies 12E7, F21, and O13
	770394	Fc fragment of IgG, receptor, transporter, alpha
	295985	ESTs
	377461	caveolin 1, caveolae protein, 22kD
BL	504791	glutathione S-transferase A4
	813742	PTK7 protein tyrosine kinase 7
	204545	ESTs
	212542	*Homo sapiens* mRNA; cDNA DKFZp586J2118 (from clone DKFZp586J2118)
	236282	Wiskott−Aldrich syndrome (ecezema-thrombocytopenia)
NB	786084	chromobox homologue 1 (*Drosophila* HP1 beta)
	810057	cold shock domain protein A
	44563	growth associated protein 43
	812105	transmembrane protein
	325182	cadherin 2, N-cadherin (neuronal)
rms	246035	ESTs
	244618	ESTs
	296448	insulin-like growth factor 2 (somatomedin A)
	796258	sarcoglycan, alpha (50 kD dystrophin-associated glycoprotein)
	784224	fibroblast growth factor receptor 4

[a] Taken from ref 13. [b] EWS, Ewing family of tumors; BL, Burkitt lymphoma; NB, neuroblastoma; rms, rhabdomyosarcoma.

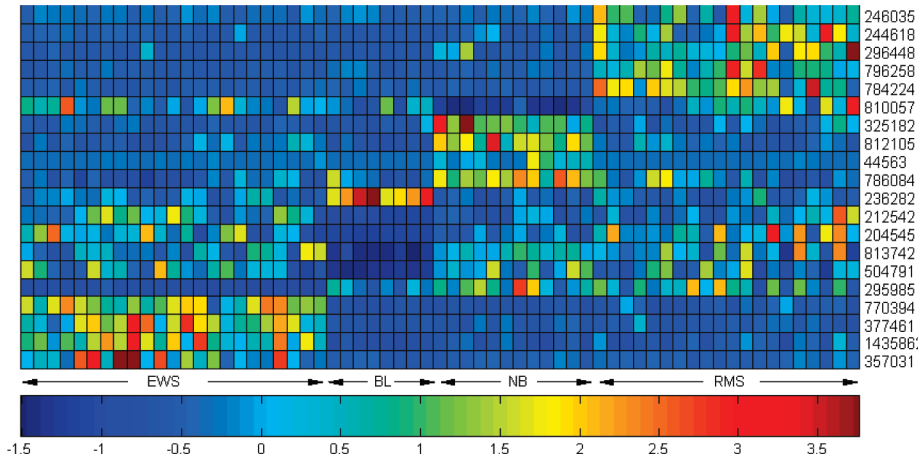

Figure 2. Heat map showing the expression levels of top five genes with the highest reselection percentages for each class in SRBCT data (20 genes in all). EWS, Ewing family of tumors; BL, Burkitt lymphoma; NB, neuroblastoma; rms, rhabdomyosarcoma. Image id on the right is taken from ref 13.

Table 3. Mean Classification Accuracy of SRBCT Data Obtained by Standard Kernel Transform-SVM or BKT-SVM Combined with Different Variable Selection Algorithms

method	no. of optimal genes selected	classification accuracy	
		training set	test set
BSS/WSS + SVM[a]	150	0.9668	0.9515
WRST + SVM[b]	40	0.9429	0.8547
ISP + SVM[c]	20	0.9820	0.9640
ISP + BKT-SVM[d]	20	1.0000	1.0000

[a] Standard kernel transform-SVM combined with BSS/WSS. [b] Standard kernel transform-SVM combined with Wilcoxon rank sum test method. [c] Standard kernel transform-SVM combined with ISP method. [d] BKT-SVM combined with ISP method.

experiment, one part of them would be used as the test set and the rest making up the training set. Then, the five different combinations of training and test sets were investigated. All procedures were all repeated five times for the five combinations. The variable selection procedures were repeated on the training set alone for each data split. The number of key genes in each model was determined by trial-and-error procedure. The average classification rates of the five computations obtained by each combined method are given in Table 3. One observed that the ISP method, followed by standard kernel transform-SVM modeling, gave the best classification rates in situations when only 20 genes were used, and the average classification rates were 98.20% and 96.40%, respectively, for the training and the test sets. However, the optimal classification was achieved using 150 genes in cases when the BSS/WSS method, followed by standard kernel transform-SVM modeling, was used but much worse average classification rates were obtained, 96.68% for the training set and 95.15% for the test set. The Wilcoxon rank sum test method followed by standard kernel transform-SVM modeling gave even worse results with average classification rates 94.29% for the training set and 85.47% for the test set, and the optimal number of selected genes was 40. Compared with the variable selection methods based on BSS/WSS and Wilcoxon rank sum test, the ISP method yielded superior classification with much less genes, indicating that the proposed strategy was much more efficient in mining the key genes with improved discriminative power. Also, the proposed BKT-SVM was employed for the

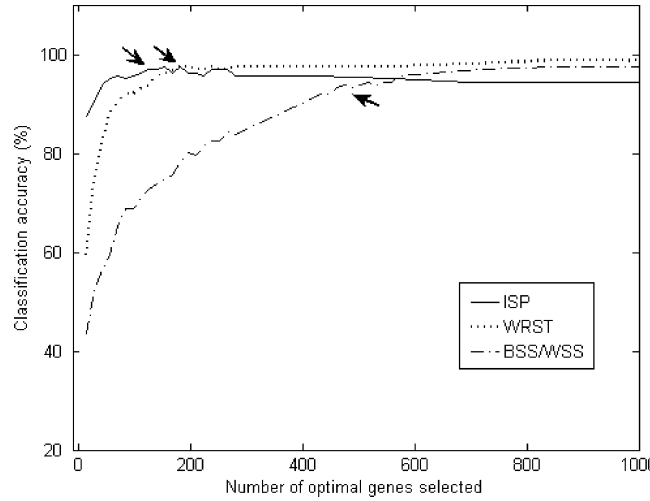

Figure 3. Performance of SVM models versus number of optimal genes selected across the three gene selection methods, ISP, WRST and BSS/WSS in GCM.

classification. The optimal number of key genes was also 20, and the average classification rate of the training and the test sets were both 100%, which was much better than the results, 98.20% and 96.40%, obtained using standard kernel transform-SVM. In contrast to standard kernel transform-SVM, the proposed BKT-SVM demonstrated improved learning and generalization ability, probably because BKT was capable of eliminating interference from genes selected for different classes.

3.2. GCM. The GCM data set[14] consists of 198 samples from 14 human tumor types represented by 16 063 genes. The 14 tumor classes were breast adenocarcinoma, prostate, lung adenocarcinoma, leukemia, colorectal adenocarcinoma, lymphoma, bladder, melanoma, uterine adenocarcinoma, renal cell carcinoma, pancreatic adenocarcinoma, ovarian adenocarcinoma, pleural mesothelioma, and central nervous system. Additional preprocessing steps were taken before standardization of the data: (1) thresholding (floor of 100 and ceiling of 16 000), (2) filtering (exclusion of genes with max/min ≤ 5 and max-min ≤ 500 across the samples), and (3) base 10 logarithmic transform and standardization.[2]

For each class the sample size was quite small, so it presented challenges for many approaches to achieve good classification. In the previous paper, nine gene selection

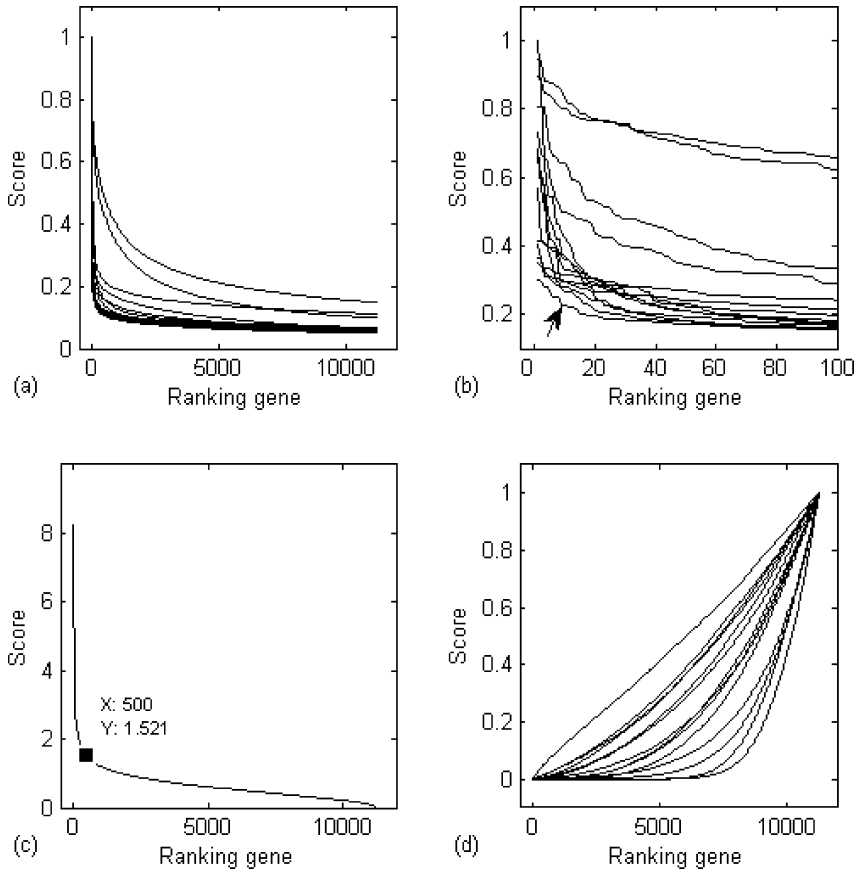

Figure 4. (a) ISP scores versus all ranked genes in each class in GCM. (b) ISP scores versus top 100 ranked genes in each class in GCM. (c) BSS/WSS scores versus all ranked genes in GCM. (d) WRST scores versus all ranked genes in each class in GCM.

methods have been employed to select the related genes followed by the treatment using six classification methods, and the best rate of correct prediction reported on GCM data set was 63.33% with more than 150 top ranked genes.[25] The best rate of correct prediction was reported to be 78.00% using one-versus-the rest SVM with all genes.[14] In the present study, 144 samples were used as the training set, and the rest, 54 samples, were taken to constitute the test set, which was consistent with that reported previously.[14] ISP method was invoked to select the key genes, and seven genes were identified for each class with desirable classification performance. That is, in total 98 genes were used for classification using BKT-SVM. A correct rate of 83.33% was obtained for the prediction set, and the classification rate was 95.83% for the training set. These results conformed again that the proposed gene selection method combined with BKT-SVM gave better results with less key genes, indicating that the proposed method could provide superior performance to other approaches in variable selection and classification.

Also, to avoid a fortuitous choice of the training and test set, the original data set was randomly divided into five completely independent parts of roughly equal size. One part of them was used as the test set and the rest making up the training set in each experiment. Then, the five different random combinations of training and test sets were investigated. Again, three gene selection methods based on BSS/WSS, Wilcoxon rank sum test and ISP were employed to identify the key genes for comparison purpose. The variable selection procedures were repeated on the training set alone for each data split. The number of key genes in each model was also determined by trial-and-error procedure. Figure 3 shows the plots of performance of the SVM models versus number of optimal genes across the three gene selection methods. It is clear that when the number of key genes is greater than 126, the variation of the performance is small for ISP. Thereby, 126 genes identified by ISP were used for a classification model. In the same way, the numbers of genes for BSS/WSS and WRST could be 182 and 500, respectively. Also, the number of genes included in a classification model combined with ISP or BSS/WSS could be determined by a plot of ISP or BSS/WSS score versus ranking gene, as shown in Figure 4a, b, and c. In Figure 4a and b, it can be observed that the ISP scores of top 9 genes in each class are obviously greater than the rest. Thus, the number of genes for ISP can be 126, which is consistent with the result revealed in Figure 3. However, it is difficult for WRST to determine the number of genes by this means, as shown in Figure 4d. The optimal genes identified by the three gene selection methods are listed in tables of the Supporting Information.

After selecting the key genes, the training and the test data sets were classified using standard kernel transform-SVM or BKT-SVM. By repeating each procedure five times for the five combinations. The statistical classification results of the five computations obtained by these algorithms, as listed in Table 4, were used to demonstrate the potential of the proposed gene selection and classification methods. One observed that with standard kernel transform-SVM procedure, the model using the genes selected by ISP method gave the best average classification rate for GCM data set, 95.70% for the training set and 83.70% for the test set. The average classification rates obtained using methods based on BSS/WSS and Wilcoxon rank sum test were rather undesirable,

Table 4. Mean Classification Accuracy of GCM Data Obtained by Standard Kernel Transform-SVM or BKT-SVM Combined with Different Variable Selection Algorithms

method	no. of optimal genes selected	mean classification accuracy	
		training set	test set
BSS/WSS + SVM[a]	500	0.9495	0.7399
WRST + SVM[b]	182	0.9432	0.7721
ISP + SVM[c]	126	0.9570	0.8370
ISP + BKT-SVM[d]	126	0.9647	0.8423

[a] Standard kernel transform-SVM combined with BSS/WSS. [b] Standard kernel transform-SVM combined with Wilcoxon rank sum test method. [c] Standard kernel transform-SVM combined with ISP method. [d] BKT-SVM combined with ISP method.

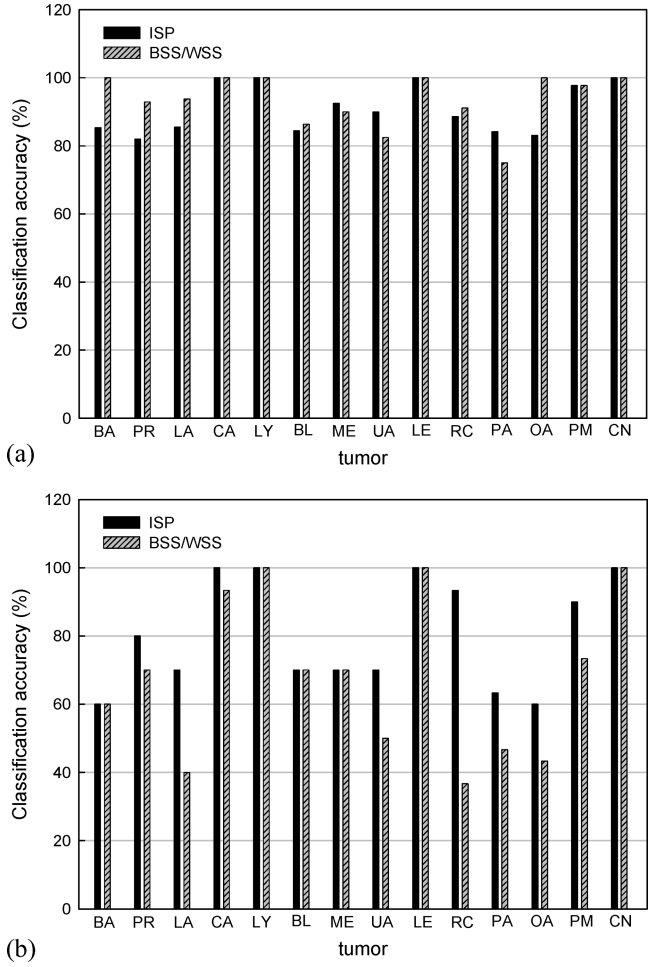

Figure 5. (a) Mean classification accuracy for each class in the training sets of GCM data obtained by the standard kernel transform-SVM combined with ISP or BSS/WSS. BA, breast adenocarcinoma; PR, prostate; LA, lung adenocarcinoma; CA, colorectal adenocarcinoma; LY, lymphoma; BL, bladder; ME, melanoma; UA, uterine adenocarcinoma; LE, leukemia; RC, renal cell carcinoma; PA, pancreatic adenocarcinoma; OA, ovarian adenocarcinoma; PM, pleural mesothelioma; CN, central nervous system. (b) The mean classification accuracy of the five computations for each class in the test sets of GCM data obtained by the standard kernel transform-SVM combined with ISP or BSS/WSS.

94.95% and 73.99% for training and test, respectively, for BSS/WSS method and 94.32% and 77.21% for training and test, respectively, for Wilcoxon rank sum test method. Figure 5 shows the average rates of correct classification and prediction of each class in GCM obtained by the standard kernel transform-SVM combined with different gene selection algorithms, BSS/WSS and ISP. From the results for the training samples illustrated in Figure 5a, it can be observed that for each class, the two methods both show high correct classification rates, implying high precision in modeling. The generalization ability that can be evaluated using the results of the test samples as illustrated in Figure 5b is very different to the two methods. For every class, ISP provides higher rates of correct prediction than BSS/WSS. Especially for prostate, lung adenocarcinoma, renal cell carcinoma, pancreatic adenocarcinoma, ovarian adenocarcinoma and pleural mesothelioma, the rates of correct prediction of 70.00%, 40.00%, 36.67%, 46.67%, 43.33%, and 73.33% obtained using BSS/WSS method were substantially improved up to 80.00%, 70.00%, 93.33%, 63.00%, 60.00%, and 90.00%, respectively, by ISP method. As a matter of fact, it was observed that the parts of the selected genes exhibited different clusters of expression levels in the samples from these diseases because of differential nosogenesis. For instance, the selected key genes of lung adenocarcinoma, SFTP3, SFTP1, and SP-C1 were found to show differential expression, which coincided with previous observations for the disease.[26] In these cases, the genes with differential expression pattern could also be identified successfully using the ISP-based gene selection method, since the method did not assume a unimodality distribution of the expression levels for each class, thereby affording better gene selection results than conventional procedures. More genes that show multimodal distribution identified by ISP but ignored by other methods are listed in table of the Supporting Information. The BKT-SVM also used to modeling using the genes selected by ISP method, and the average classification rates were 96.28% for the training sets and 84.23% for the test sets. Compared with standard kernel transform-SVM, BKT-SVM again showed improved classification performance because of the incorporation of BKT.

4. CONCLUSION

A new gene selection method based on interval segmentation purity, combined with a block-wise kernel transform based support vector machine, was proposed as an efficient approach for combating the challenge of high-dimensional microarray data with redundancy of variables and scarcity of samples. This new strategy furnished a unique advantage over existing methods in the capability of selecting discriminative genes with multimodal expression patterns, which coincided with the common situations where a single disease state had multiple subtypes derived from varying pathogenic mechanisms. The results demonstrated that the developed gene selection method could identify a key gene set with the least size but the optimal classification performance. Furthermore, the block-wise kernel transform provided superior classification than standard one due to its capability of eliminating interference from redundant genes from different classes. Therefore, the proposed approach was expected to hold great promise for microarray data analysis in gene marker discovery and expression profile-based clinical diagnosis.

ACKNOWLEDGMENT

This work was supported by "973" National Key Basic Research Program (2007CB310500) and National Nature

Science Foundation (Grant No. 20775023, 20675028, J0830415) of China.

Supporting Information Available: Detailed structural formulas of the compounds. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

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

(2) Dudoit, S.; Fridlyand, J.; Speed, T. P. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **2002**, *97*, 77–87.

(3) Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G. Class prediction by nearest shrunken centroids, with application to DNA microarrays. *Stat. Sci.* **2003**, *18*, 104–117.

(4) Boulesteix, A. L.; Tutz1, G.; Strimmer, K. A. CART-based approach to discover emerging patterns in microarray data. *Bioinformatics* **2003**, *19*, 2465–2472.

(5) Sato, H.; Grutters, J. C.; Pantelidis, P.; Mizzon, A. N.; Ahmad, T.; Houte, A. J. V.; Lammers, J. W. J.; Bosch, J. M. M.; Welsh, K. I.; Bois, R. M. HLA-DQB1*0201: a marker for good prognosis in british and dutch patients with sarcoidosis. *Am. J. Respir. Cell Mol. Biol.* **2002**, *27*, 406–412.

(6) Katahira, J.; Sugiyama, H.; Inoue, N.; Horiguchi, Y.; Matsuda, M.; Sugimoto, N. *Clostridium perfringens* enterotoxin utilizes two structurally related membrane proteins as functional receptors in vivo. *J. Biol. Chem.* **1997**, *272*, 26652–26658.

(7) Cheng, Y. Z. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Machine Intell.* **1995**, *17*, 790–799.

(8) Fukunaga, K.; Hostetler, L. D. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theor.* **1975**, *21*, 32–40.

(9) Cheng, Y.; Fu, K. S. Conceptual clustering in knowledge organization. *IEEE Trans. Pattern Anal. Machine Intell.* **1985**, *7*, 592–598.

(10) Vapnik, V. N. In *The Nature of Statistical Learning Theory*; Springer: New York, 1995.

(11) Vapnik, V. N. *Statistical Learning Theory*; Wiley: New York, 1998.

(12) Du, Y. P.; Liang, Y. Z.; Li, B. Y.; Xu, C. J. Orthogonalization of block variables by subspace-projection for quantitative structure property relationship (QSPR) research. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 993–1003.

(13) Khan, J.; Wei, J. S.; Ringnér, M.; Saal, L. H.; Ladanyi, M.; Westermann, F.; Berthold, F.; Schwab, M.; Antonescu, C. R.; Peterson, C.; Meltzer, P. S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* **2001**, *7*, 673–679.

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

(15) Okada, K.; Singh, M.; Ramesh, V. Prior-constrained scale-space mean shift. In *BMVC'2006*; Proceedings of the 17th British Machine Vision Conference, Edinburgh, U.K., 4−7 September 2006; BMVA Press: Bristol, 2006; pp 829−838.

(16) Lin, W. Q.; Jiang, J. H.; Zhou, Y. P.; Wu, H. L.; Shen, G. L.; Yu, R. Q. Support vector machine based training of multilayer feedforward neural networks as optimized by particle swarm algorithm: Application in QSAR studies of bioactivity of organic compounds. *J. Comput. Chem.* **2007**, *28*, 519–527.

(17) Tang, L. J.; Zhou, Y. P.; Jiang, J. H.; Zou, H. Y.; Wu, H. L.; Shen, G. L.; Yu, R. Q. Radial basis function network-based transform for a nonlinear support vector machine as optimized by a particle swarm optimization algorithm with application to QSAR studies. *J. Chem. Inf. Model.* **2007**, *47*, 1438–1445.

(18) Benoudjit, N.; Archambeau, C.; Lendasse, A.; Lee, J.; Verleysen, M. Width optimization of the gaussian kernels in radial basis function networks. In *ESANN'2002 Proceedings*; European Symposium on Artificial Neural Networks, Bruges, Belgium, 24−26 April 2002; D-Side: Bruges, 2002; pp 425−432.

(19) Wan, X.; Helman, L. J. Effect of insulin-like growth factor II on protecting myoblast cells against cisplatin-induced apoptosis through p70 S6 kinase pathway. *Neoplasia* **2002**, *4*, 400–408.

(20) Xie, C.; Lovell, M. A.; Xiong, S.; Kindy, M. S.; Guo, J.; Xie, J.; Amaranth, V.; Montine, T. J.; Markesbery, W. R. Expression of glutathione-S-transferase isozyme in the SY5Y neuroblastoma cell line increases resistance to oxidative stress. *Free Radical Biol. Med.* **2001**, *31*, 73–81.

(21) Girnita, L.; Girnita1, A.; Wang, M.; Meis-Kindblom, J. M.; Kindblom, L. G.; Larsson, O. A link between basic fibroblast growth factor (bFGF) and EWS/FLI-1 in Ewing's sarcoma cells. *Oncogene* **2000**, *19*, 4298–4301.

(22) Pasic, S.; Vujic, D.; Djuricic, S.; Jevtic, D.; Grujic, B. Burkitt lymphoma-induced ileocolic intussusception in Wiskott−Aldrich syndrome. *J. Pediatr. Hematol./Oncol.* **2006**, *28*, 48–49.

(23) Minniti, C. P.; Luan, D.; Grady, C. O.; Rosenfeld, R. G.; Oh, Y.; Helman, L. J. Insulin-like growth factor II overexpression in myoblasts induces phenotypic changes typical of the malignant phenotype. *Cell Growth Differ.* **1995**, *6*, 263–269.

(24) Hollander, M.; Wolfe, D. A. In *Nonparametric Statistical Inference*; Wiley: New York, 1973.

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

(26) Dutu, T.; Michiels, S.; Fouret, P.; Penault-Llorca, F.; Validire, P.; Benhamou, S.; Taranchon, E.; Morat, T.; Grunenwald, D.; Chevalier, T. L.; Sabatier, L.; Soria, J. C. Differential expression of biomarkers in lung adenocarcinoma: a comparative study between smokers and never-smokers. *Ann. Oncol.* **2005**, *16*, 1906–1914.

CI900032Q