Cancer Classification from Gene Expression Data by NPPC Ensemble

Santanu Ghorai, Anirban Mukherjee, Sanghamitra Sengupta, and Pranab K. Dutta

Abstract—The most important application of microarray in gene expression analysis is to classify the unknown tissue samples according to their gene expression levels with the help of known sample expression levels. In this paper, we present a nonparallel plane proximal classifier (NPPC) ensemble that ensures high classification accuracy of test samples in a computer-aided diagnosis (CAD) framework than that of a single NPPC model. For each data set only, a few genes are selected by using a mutual information criterion. Then a genetic algorithm-based simultaneous feature and model selection scheme is used to train a number of NPPC expert models in multiple subspaces by maximizing cross-validation accuracy. The members of the ensemble are selected by the performance of the trained models on a validation set. Besides the usual majority voting method, we have introduced minimum average proximity-based decision combiner for NPPC ensemble. The effectiveness of the NPPC ensemble and the proposed new approach of combining decisions for cancer diagnosis are studied and compared with support vector machine (SVM) classifier in a similar framework. Experimental results on cancer data sets show that the NPPC ensemble offers comparable testing accuracy to that of SVM ensemble with reduced training time on average.

Index Terms—Cancer classification, classifier ensemble, combination of multiple classifiers, microarray data analysis, proximal classifier.

1 Introduction

The microarray technique has led the scientists immense opportunity to measure the expression levels of thousands of genes simultaneously in a single experiment [1], [2], [3]. The advantage of this method, namely to monitor a large number of variables of a sample of tissue's state, however, often turns out to be difficult to analyze. The most important application of the microarray technique is to classify unknown samples according to their expression profile, e.g., to discriminate cancerous or noncancerous samples or to discriminate different types or subtypes of cancer [4], [5], [6], [7], [8], [9]. The small number of samples and the level of noise make the classification task of a test sample challenging. To accomplish this, the first step of processing the expression data is to identify a small subset of genes that are primarily responsible for the disease [10], [11], [12], [13]. This will serve the purpose of looking deep insight into the nature of the disease, genetic mechanism responsible for it [14], drug discovery for the disease [15], [16], [17]. The small subset of genes will also provide improved diagnostic accuracy and reduce the cost of microarray array experiment by reducing the chip size, manpower, and easier interpretable experiments [18].

- *S. Ghorai is with the Department of Electronics and Communication Engineering, MCKV Institute of Engineering, 243, G.T. Road (N), Liluah, Howrah 711204. E-mail: san_ghorai@yahoo.co.in.*
- *A. Mukherjee and P.K. Dutta are with the Department of Electrical Engineering, Indian Institute of Technology, Kharagpur 721302, West Bengal, India. E-mail: {anirban, pkd}@ee.iitkgp.ernet.in.*
- *S. Sengupta is with the Department of Biochemistry, University of Calcutta, 35, Ballygunge Circular Road, Kolkata 700 019, West Bengal, India. E-mail: sanghamitrasg@yahoo.com.*

Manuscript received 26 June 2009; revised 16 Oct. 2009; accepted 26 Oct. 2009; published online 30 Apr. 2010.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2009-06-0107. Digital Object Identifier no. 10.1109/TCBB.2010.36.

In the literature, there are several methods of feature or gene selection. All these methods can be divided into three categories: filter methods, wrapper methods, and embedded methods [19]. The filter methods are used to extract those features which show dependences on the class labels without explicitly relying on a classifier. Examples are methods based on statistical ranking of individual genes, such as, correlation coefficient [20], t-statistics [21], [22], class seperability [23], or Fisher's criterion, etc. [24], [25]. Additionally, there are methods which consider the mutual information among the genes as well as the relevance of the genes for classification [26]. All these methods are fast compared to other two methods, i.e., wrapper and embedded methods. The wrapper methods [27], [28], [29] use a classifier as the objective function for the evaluation of a subset of features. The classifier is obtained by a model selection method which maximizes the classification accuracy on a validation set. This validation set is kept separated from the training data. Typical classifiers used for this purpose are Bayesian classifier [30], [31], K-nearest neighbor [31], [32], support vector machine (SVM) [10], [33], relevance vector machine [34], penalized kernel logistic regression (PKLR) classifier [35], etc. The wrapper methods are very slow as they search several combinations of genes and optimal parameter set. In embedded method, the genes are selected as part of the specific learning method. Examples are one-norm SVM [36], logistic regression [37], sparse logistic regression [38], [39], probit regression [40], joint classifier and feature optimization (JCFO) [41], methods based on regularization technique [42], etc.

All the above variations provide comparable feature selection and classification accuracy. The most important fact in a medical diagnosis system is the classification accuracy of unknown samples (generalization performance).

To improve the classification accuracy, the gene selection should be a part of the learning procedure. But the large number of classifiers required to build (train and test) makes this method very slow. On the other hand, the genes selected by the filter method may not be the optimal feature set for a classifier to obtain high classification accuracy. Moreover, the model parameter selection is another crucial point to obtain high classification performance. Besides all these, several research works were performed on selection of optimal number of genes that are sufficient to classify a data set accurately. Again, most of the papers reported cross-validation (CV) testing accuracy of their methods which suffers from the "selection bias" as the testing sample is not excluded from the gene selection procedure [43]. In order to evaluate the true performance of a computer-aided diagnosis (CAD) method, it is mandatory to exclude the testing samples from the classifier building process, i.e., data normalization, gene selection, and model parameter selection [43], [44].

In this paper, we have proposed a hybrid CAD method of cancer classification taking the advantage of both filter and wrapper methods. A fast dimensionality reduction step is carried out by selecting a small set of genes by the MRMR [26] ranking method. Then we used the wrapper method (for improved classification accuracy) on this small set of genes to reduce the computational burden. We have selected nonparallel plane proximal classifier (NPPC) [45], [46] as a part of the wrapper method. In our previous research work, we have proposed NPPC [45], [46] for binary data classification that provides comparable accuracy with that of SVM classifiers [47], [48], [49] with a lower computational cost. In [45], we have focused on the chronological development of NPPC having its root from SVM. For a binary data classification problem, SVM finds an optimal hyperplane that maximizes the separation between the two classes in $O(m^3)$ iterations, where m is the number of training data. The least-squares SVM (LS-SVM) finds the same optimal hyperplane in less than $O(m^3)$ iterations. The proximal SVM (PSVM) finds two parallel planes in $O(n^3)$ iterations, where $n(\ll m)$ is the number of features. PSVM classifies the data according to the proximity to these two planes. The generalized eigenvalue proximal SVM (GEPSVM) [50] finds two nonparallel planes such that the first plane is "closest" to the positive examples and "furthest" from the negative examples and the second plane has the opposite properties. It then classifies according to proximity to these two nonparallel planes. The twin SVM (TWSVM) [51] is an alternative formulation of GEPSVM which can be solved in $O(m^3/4)$ training iterations given that positive and negative data sets have approximately equal cardinality. NPPC combines ideas from the TWSVM and PSVM. NPPC generates two unity norm nonparallel planes [46] by solving two equality-constrained (like PSVM) optimization problems. This idea of nonparallel plane classifier differs from that of the classical SVM, which is based on the margin maximization of two separating parallel hyperplanes [47], [48]. But NPPC is computationally more efficient than SVM classifier [46]. The equality constraints accomplish the training task of the NPPC faster than its inequality counterpart in TWSVM. On the other hand, it is experimentally observed that the relaxation of the parallelism constraint of the separating planes offers comparable classification accuracy on the noisy data sets as compared to SVM [45], [46].

The excellent performance of NPPC motivated us to apply it on microarray data, which is inherently noisy, for discrimination of cancerous and normal tissue samples. But we have observed that the classification accuracy of a single NPPC is not satisfactory on microarray data by using a small set of informative genes selected by the filter method. To improve the diagnostic accuracy, we have introduced NPPC ensemble (in place of a single NPPC) that consists of a number of experts trained with the best gene subset of different cardinality. The classifier ensemble or committee is a widely explored topic in machine learning applications [52], [53], [54], [55]. Multiple classifiers are combined with the expectation that it will perform better than (at least same as) a single classifier. But the selection of optimal classifiers in the ensemble [55, ch. 6], [56], [57], [58] and their combination rule [59], [60] are still an active area of research in the machine learning community. Additionally, the model selection task of NPPC by a grid search method is computationally intensive as it has four regularization parameters for a linear classifier. As a result, the realization of a number of best NPPC models with all possible gene subsets to form the ensemble will be computationally unrealistic. To overcome this drawback, we have introduced a genetic algorithm (GA) [61] based simultaneous feature (gene) and model parameter selection scheme to train a number of experts with different cardinality. The NPPC ensemble is formed by selecting the trained expert models based on the performance on a validation set. Besides the well-known majority voting scheme of decision combination of ensemble, we have proposed a new decision combination schemes for NPPC ensemble based on the proximity profile of each test pattern. We have compared our method with SVM classifier. Experimental results show that the GA-based optimal model selection scheme is computationally efficient than that of the conventional grid search method. It has also been observed that both NPPC and SVM ensembles perform better than the respective single classifier. From the experimental results on benchmark microarray data sets, it is evident that both the NPPC and SVM ensembles offer high classification accuracy, but the training time for the NPPC ensemble is reduced by 40-80 percent compared to that of the SVM ensemble. It is to be noted that the NPPC ensemble does not address the small sample problem of data sets. It only offers improved classification performance on noisy microarray data sets.

This paper is organized as follows: in Section 2, we briefly describe the NPPC formulation. In Section 3, we describe the development methodology of NPPC ensemble. Additionally, we have introduced a new proximity profile-based decision combiner for multiple classifiers. The experimental results are presented in Section 4 and discussed in Section 5. Finally, Section 6 concludes the paper.

2 NPPC Formulation

2.1 Basic Formulation for Binary Data Classification

In [46], we have formulated NPPC with unity norm hyperplanes that overcome the drawback of classical NPPC [45]. NPPC is a nonparallel plane classifier that classifies binary data by its proximity to one of the two nonparallel

planes. The two planes are obtained by solving two nonlinear programming problems (NPPs) with a quadratic form of the loss function. Each plane is clustered around a particular class of data by minimizing the sum square distances of patterns from it and considering the patterns of the others class at a euclidean distance of 1 with errors. Thus, the objective of NPPC is to find two hyperplanes

$$\omega_1^T x + b_1 = 0 \quad \text{and} \quad \omega_2^T x + b_2 = 0, \tag{1}$$

where $\omega_1, \omega_2 \in \Re^n$ and $b_1, b_2 \in \Re$ are normal vectors and bias terms of the hyperplanes 1 and 2, respectively. To obtain the above two planes, NPPC solves the following pair of NPPs:

$$\begin{aligned} \underset{(\omega_1, b_1, \xi_2) \in \Re^{(n+1+m_2)}}{\text{Min}} \; & J_1(\omega_1, b_1, \xi_2) \\ & = \frac{1}{2}\|A\omega_1 + e_1 b_1\|^2 + c_1 e_2^T \xi_2 + \frac{c_2}{2}\|\xi_2\|^2 \qquad (2) \\ \text{s.t.} \quad & -(B\omega_1 + e_2 b_1) + \xi_2 = e_2 \\ \text{and} \quad & \|\omega_1\| = 1, \end{aligned}$$

$$\begin{aligned} \underset{(\omega_2, b_2, \xi_1) \in \Re^{(n+1+m_1)}}{\text{Min}} \; & J_2(\omega_2, b_2, \xi_1) \\ & = \frac{1}{2}\|B\omega_2 + e_2 b_2\|^2 + c_3 e_1^T \xi_1 + \frac{c_4}{2}\|\xi_1\|^2 \qquad (3) \\ \text{and} \quad \text{s.t.} \quad & (A\omega_2 + e_1 b_2) + \xi_1 = e_1 \\ \text{and} \quad & \|\omega_2\| = 1, \end{aligned}$$

where matrices $A \in \Re^{m_1 \times n}$ and $B \in \Re^{m_2 \times n}$ contain the m_1 and m_2 training patterns of class 1 and class -1, respectively, in n-dimensional space; $e_1 \in \Re^{m_1}$ and $e_2 \in \Re^{m_2}$ are vectors of ones and $\xi_1 \in \Re^{m_1}, \xi_2 \in \Re^{m_2}$ are error variable vectors due to class 1 and class-1 data, respectively; and $c_1, c_2, c_3,$ and $c_4 > 0$ are four regularization parameters of the NPPC. The first term of each objective function minimizes the sum of the squared distances from the hyperplane to the patterns of respective class and the constraint requires that the patterns of opposite class are at a euclidean distance of 1 from the hyperplane with errors. The second and third terms of the objective functions constitute the general quadratic loss function. The inclusion of the constraints makes the problem nonlinear but reduces it to a convex optimization problem of least-squares type. We have replaced the original problems (2) and (3) by a penalty function [62, pp. 497-527] approach and then employed the quadratically convergent Newton's method [62, pp. 44-49] with Armijo steps [63], [64] to solve the problem in finite number of iterations. The advantage of the NPPC is that its training can be accomplished by solving two-system of linear equations, instead of solving a quadratic program as it requires for training standard SVM classifiers [47], [48], [49]. For linear NPPC, it solves two-system of linear equations in n-dimensional space, where n is the number of attributes. Thus, the problem can be solved very fast by using many sophisticated algorithms that are used to solve system of linear equations. Here, we have used the conjugate gradient method to solve the system of linear equations in each iteration. Readers may see [46] for a detail solution and implementation of NPPC. Once the training of the classifier is accomplished, a new data sample $x \in \Re^n$ is assigned to a class l by comparing the following distance measure of it from the two hyperplanes given by (1), i.e.:

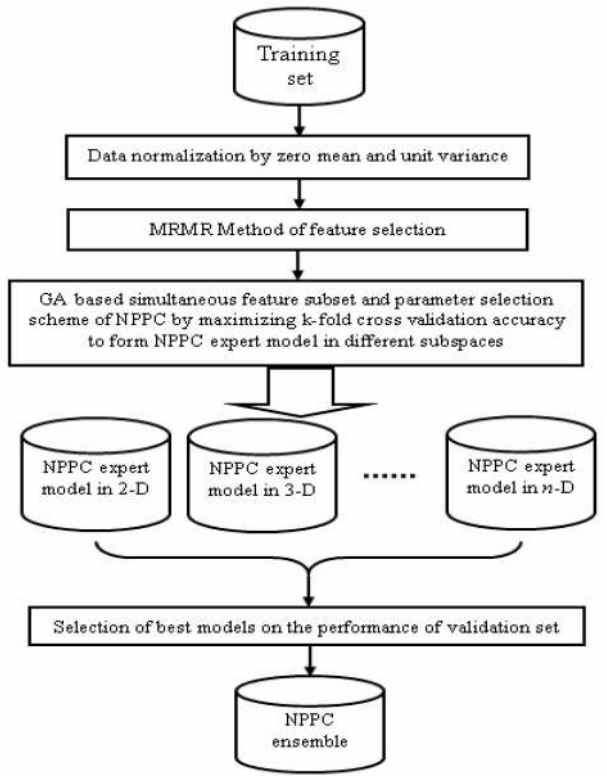

Fig. 1. Block diagram showing the steps of formulation of NPPC ensemble.

$$Class\ l = \arg\left(\underset{r=1,2}{Min} \|\omega_r^T x + b_r\| \right). \tag{4}$$

3 NPPC Ensemble Formulation

In this section, we have discussed the various steps of formulation of NPPC ensemble. As we have pointed out earlier that the computer-aided diagnosis of cancer or cancer category discrimination should be bias-free, we have preceded the formulation of NPPC ensemble by separate training, validation, and testing set. The validation and testing set is normalized by subtracting the mean and dividing the standard deviation of the training data. The schematic block diagram of NPPC formulation is shown in Fig. 1.

The training module of our proposed NPPC ensemble technique consists of four steps.

1. Gene selection by the filter method,
2. formation of NPPC experts in different dimensions by the wrapper method,
3. selection of expert models in the NPPC ensemble by the performance of trained classifiers on validation set, and
4. decision combination of the selected NPPC experts in the ensemble. These steps are discussed as follows:

3.1 Gene Selection by the Filter Method

The earlier filter-based methods, such as correlation-based methods [20], t-statistics [21], [22], or F-statistics [23], [24], [25], operate in isolation for ranking the genes and do not consider the correlation among the features. Thus, the

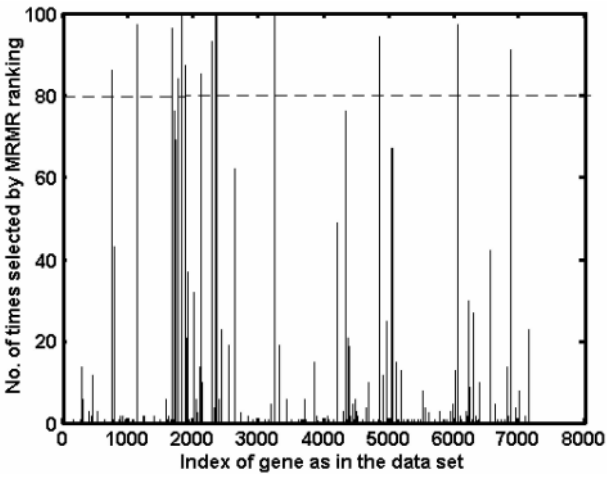

Fig. 2. Frequency of genes selected in the ranking by the MRMR method for ALL_AML data set.

redundancy among the selected features is not used. To overcome this difficulty, Ding and Peng [26] proposed the minimum redundancy maximum relevance (MRMR) method of feature selection that utilizes the mutual information criteria for selecting a set of most informative features. They take into account the maximum relevance along with the minimum redundancy criteria to choose the additional features that are maximally dissimilar to the already identified features.

We have randomly divided the available samples into training and testing sets with a sample ratio 6:4. Then, the MRMR method with the quotient scheme is applied to the training data to select the highest ranked 25 genes. This experiment is performed 100 times for each data set with random permutation. We have observed that only few particular genes repeatedly selected in the highest ranking of 25 gene subset. We have chosen only those genes for a data set which are selected at least 80 times (out of 100 test runs) in the list of 25 highest ranked candidates. As an example, the frequency distribution of selected genes in the subset (25 in number) of MRMR ranking obtained in the experiment for ALL_AML data set [4] is shown in Fig. 2. This shows that only 13 genes hit more than 80 times in the ranking. We have selected these 13 genes to construct the NPPC ensemble. Thus, we preselected a few genes, typically less than 15, using the filter method. This reduced set of genes is used to construct NPPC ensemble in the next stage. The lists of selected genes by this method for different data sets are provided as supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2010.36) and discussed in Section 5.

3.2 Formation of NPPC Experts by GA

Once the gene subset is selected by the filter method containing n ($= 13$ for ALL_AML data set) best genes, we frame $(n-1)$ number of NPPC expert models, each of dimensions i with $2 \leq i \leq n$, by the wrapper method. We have excluded the one-dimensional classifier model since classification by one feature usually requires a threshold which is beyond the scope of this NPPC. Since NPPC

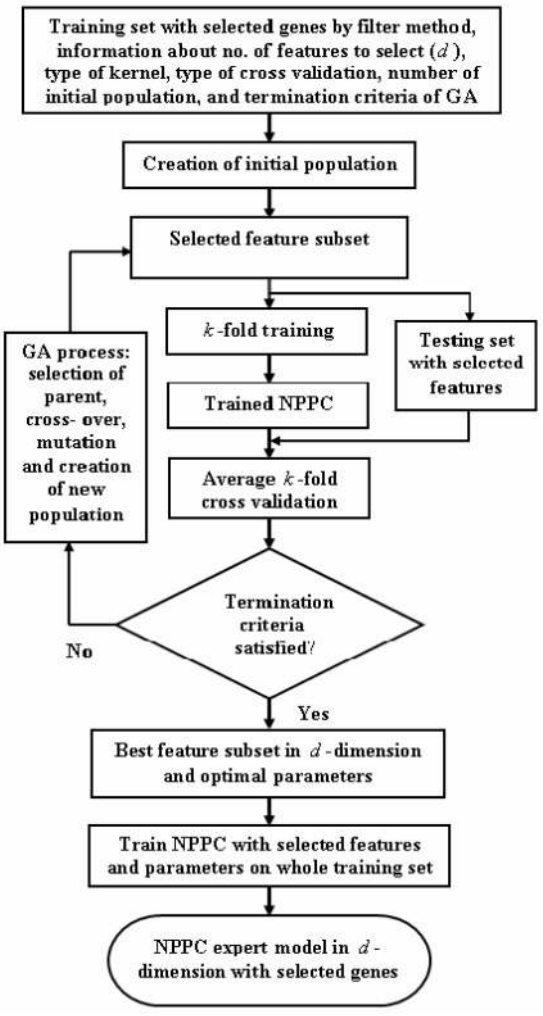

Fig. 3. System architecture of the proposed GA-based simultaneous feature and model selection.

classifies by two nonparallel planes with unity norm, the classifier in one dimension chooses the same hyperplanes with $\omega_1 = 1$ and $\omega_2 = 1$. Thus, we have intended at searching the discrimination ability of NPPC in different subspaces with the best gene subset $\{s_i\}$ in the ith dimension for $i = 2, 3, \ldots, n$. For this purpose, we have used GA [61] to simultaneously choose the best set of genes and the four tuning parameters c_1, c_2, c_3, and c_4 of the NPPC by maximizing the k-fold cross-validation accuracy [65]. GA has the potential to simultaneously generate both optimal feature subset and tuning parameters [66], [67]. For the selection of NPPC expert models in any subspace, we have used the system architecture as shown in Fig. 3. The authors of [28] used a similar type model selection scheme for SVM.

As shown in Fig. 3, the algorithm starts with the various initial values of GA parameters and the number of features to select for developing a classifier. For linear kernel NPPC, the chromosome comprises five parts: c_1, c_2, c_3, c_4, and the feature musk of size n. The structure of a chromosome for linear NPPC is shown in Fig. 4. Each parameter is represented by a gene of r bits, which can be selected according to the calculation precision required. The minimum and maximum values of the parameters are decided by the users. Feature musk consists of n-bits. A value of 1 in

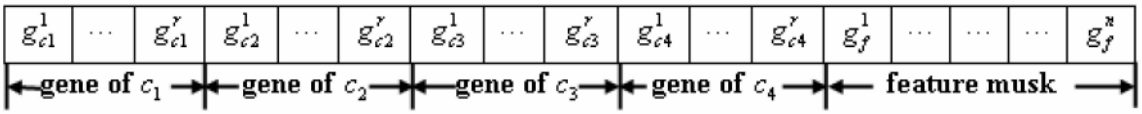

Fig. 4. Structure of chromosome for simultaneously selecting optimal features and parameters.

any bit selects the corresponding feature and a value of 0 disables the same. To develop a model in a d-dimensional subspace ($d \in \{2, 3, \ldots, 13\}$ for ALL_AML data set), only d number of bits are set at 1 out of n-bits of the feature mask. For each chromosome representing the parameters and selected features, k-fold CV testing accuracy on the training data set is selected as the fitness function to be maximized. Depending upon the fitness value of the chromosomes, it selects the parents for crossover, mutation, etc. to generate a new population. After the termination of the GA process, we trained the final NPPC model in d-dimension with the selected d features and four optimal tuning parameters c_1, c_2, c_3, c_4.

3.3 Selection of Expert Models in the NPPC Ensemble

The idea of combining classifiers in an ensemble or committee is based on the expectation that the committee can take better decision than that of its individual member. This can be true if the effective member classifiers are selected to form the classifier ensemble. Several researchers have focused on different methods for combining multiple classifiers [52], [55], [56], [57]. The formulation of NPPC ensemble is described as follows: we have a set of $(n-1)$ number of trained expert models of dimensions $2, 3, \ldots, n$. Among these expert models, we have to choose $Z (Z \leq (n-1))$ members such that their combined decisions reduce the generalization error.

Among the several methods of classifier selection, it is difficult to choose a particular method which will outperform others for all applications. Here, in this work, we adopt a simple method of classifier selection (as a member of ensemble) by their performance on the validation data. Once the expert models are trained for different dimensions, we test their performance on a validation set to choose the best experts to be included in the ensemble. The validation set is normalized by subtracting the mean and dividing the standard deviation of the training data. Only a few expert models are selected to form the ensemble which exceeds a predefined threshold p_{th} (in percentage) of the validation accuracy.

3.4 Decision Combination of the Selected NPPC Experts in the Ensemble

The research focus of combination of multiple experts (CME) is to find a combination function $f(E_1, E_2, \ldots, E_Z)$ that will produce the best identity of a new test pattern, where $E_i, i = 1, 2, \ldots, Z$, is the ith expert classifier. The possible ways of combining the outputs of Z experts in an ensemble depend on the information available from the individual member expert E_i. There are many CME-based approaches, among which majority voting, weighted majority voting, Bayes combination, behavior knowledge space methods, Dempster-Shafer (DS) combination methods, etc. are the most representative [55, ch. 4, and 5] combination methods.

In this work, we have studied two combination schemes for NPPC ensemble. The first one is the majority voting method. The other one, we have proposed, is the minimum average proximity of a pattern for combining decisions of NPPC ensemble. Since NPPC classifies a pattern by its proximity to one of the two hyperplanes, we exploit this measurement level output of NPPC for decision combination. Let $x \in \Re^n$ be a test pattern in n-dimensional subspace. NPPC assigns a class label to it by comparing its distance from the two hyperplanes. Thus, instead of deriving the absolute class labels from each expert E_i, we have calculated the proximity profile (in absolute sense) of the test pattern from the hyperplanes. Different experts derive the proximity of the test pattern in different subspaces. So, each expert $E_i \in E$ in the ensemble $E = \{E_1, E_2, \ldots, E_Z\}$ outputs two proximity values of a test pattern. The smaller the proximity, more likely is the class label. The output of $E = \{E_1, E_2, \ldots, E_Z\}$ for a particular test pattern $x \in \Re^n$ can be organized in a proximity profile matrix ($PP(x) \in \Re^{Z \times 2}$) as shown in Fig. 5. The entries in columns 1 and 2 of $PP(x)$ are individual proximity figures for class labels 1 and 2, respectively. Thus, each column develops an overall proximity vector d_j, $j = 1, 2$, for the jth class obtained from different subspace experts E_i of the ensemble $E = \{E_1, E_2, \ldots, E_Z\}$. Finally, the output class of a test pattern x by this method is decided by the minimum of average column-sum of $PP(x)$, i.e.,

$$cl(x) = \arg \min_{j=1}^{2} \left(\sum_{i=1}^{Z} d_{ij}/Z \right). \tag{5}$$

The testing of a new pattern by NPPC ensemble is summarized in Fig. 6. To test a new pattern, the preselected (by MRMR during training) n number of genes are chosen. Then this chosen n-dimensional vector is normalized by subtracting the mean and dividing the standard deviation of the training data.

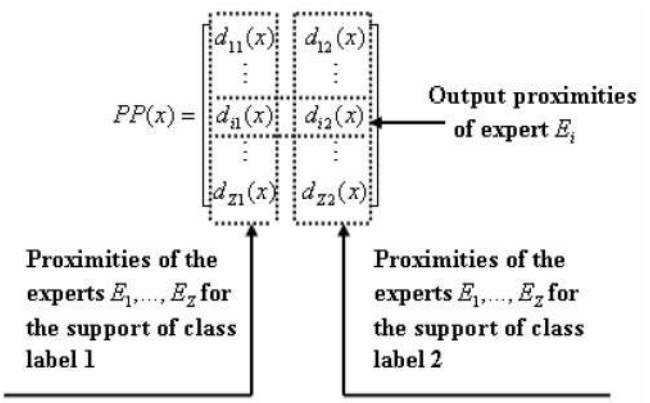

Fig. 5. Proximity profile (PP) matrix of the test pattern x.

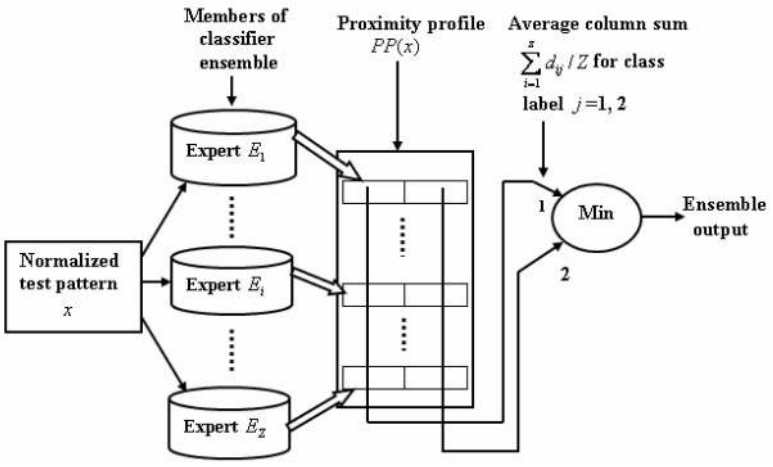

Fig. 6. Decision combination of the NPPC ensemble by the minimum average proximity of a test pattern.

4 Experiments

We have conducted numerous assessments of the proposed NPPC ensemble for the discrimination of cancer and normal tissue samples. We provided a comprehensive study on the performance of our method and compared the results with SVM in a similar framework. After a brief introduction of the data sets in this section, we have described the design of experiment followed by the comparison of performances.

4.1 Data Sets

We have selected seven public domain microarray data sets for experimental study. The details of the data sets, e.g., their availability, number of samples, number of genes, and partition for the experiment, are given in Table 1. We have omitted the detailed discussion and sources of the data sets, as these are described in [44] and [71]. A *k*-nearest neighbor algorithm [72] has been applied to fill the missing values of the Lymphoma and Liver cancer data sets. To honestly evaluate the performance of NPPC ensemble, we have separated the training and testing samples at first. We have divided the available samples randomly into training and testing in a ratio of 6:4. Further, we separate the validation set from the training set by taking 20 percent samples from the training set. All the data set is normalized (featurewise) by subtracting the training mean and then dividing by the standard deviation of the same training data.

4.2 Experiment Design

In order to test the performance of the NPPC ensemble, we have implemented it in MATLAB 7 [73] in Windows XP running on a PC with system configuration Intel P4 processor (3.06 GHz) with 1 GB of RAM. To compare the performances of the NPPC classifier with SVM in a similar programming environment, we have used the Gunn SVM toolbox [74] implemented in MATLAB. We have performed all the experiments with linear classifiers only. To honestly evaluate the performance of the NPPC ensemble, we have utilized the strategy to select genes only from the training sample. The testing and the validation samples are totally excluded from the classifier building process, e.g., data normalization, gene selection and simultaneous feature subset, and parameter selection by GA. In all the experiments, the search regions of the regularization parameters of NPPC are taken as [0 10.0] for c_1 and c_3 and $[10^{-8}\ 10]$ for c_2 and c_4. For SVM, the regularization parameter is selected in the range $[10^{-8}\ 500]$. To construct the linear NPPC ensemble for each data set in different subspaces, we have initialized the parameters randomly within the above limits and considered initial population of 20. We set the GA to run for 100 iterations. The iteration stops if the fitness value remains constant for $\tau_{gen} = 25$ generations or reaches at the maximum cross-validation accuracy of 100. The best solution at each generation is updated if the minimum change of fitness value is 10^{-6}. As the fitness function, we have considered 10-fold cross-validation (CV10) accuracy of

TABLE 1
Cancer Data Sets Used for the Experimental Study

Name	Reference	No. of samples	No. of genes	No. of genes selected by MRMR	No. of samples used for experiment in the		
					Training set	Validation set	Testing set
ALL-AML	[4]	72	7129	13	32	8	32
Colon cancer	[68]	62	2000	14	31	7	24
Lung cancer	[6]	181	12533	10	88	21	72
Breast cancer (ER)	[5]	49	7129	9	24	6	19
Lymphoma	[69]	77	7129	9	38	9	30
Liver cancer	[70]	156	1648	8	76	18	62
Prostate cancer	[9]	102	12600	9	50	12	40

TABLE 2
Performance Comparison between Grid Search Technique and GA-Based Model Selection of NPPC

Data set	Model selection by grid search technique		Model selection by GA	
	Average CPU time (Sec.)	Average % testing accuracy	Average CPU time (Sec.)	Average % testing accuracy / p-value
ALL-AML	584.47	87.08±9.67	**52.82**	**94.52±4.56 0.000102**
Colon cancer	629.84	74.44±8.75	**109.25**	**82.77±4.946 0.005011**
Lung cancer	688.68	94.35±3.21	**45.46**	96.38±3.27 0.226078
Breast cancer (ER)	587.26	67.72±9.58	**48.16**	**81.21±8.75 6.8402E-8**
Lymphoma	692.53	83.55±7.18	**57.64**	86.88±6.25 0.131848
Liver cancer	678.74	91.82±5.19	**57.76**	**96.77±3.71 0.022223**
Prostate cancer	627.38	79.5±10.91	**92.42**	**90.16±8.74 0.001847**

training data. CV10 accuracy is preferred to leave-one-out cross-validation (LOOCV), as the former has lower variance. Previous study has shown that the CV10 is more appropriate when considering the compromise between bias and variance [43], [65]. Moreover, CV10 is computationally lucrative than that of LOOCV. We have selected only those models in the ensemble for which the validation accuracy is greater than 80 percent. In order to assess the statistical significance of the proposed ensemble method, we repeated the experiment for each data set 30 times with the same selected few genes by the MRMR method as discussed in Section 3. The average test error rates and their standard deviations over the 30 experiments are reported for each data set.

4.3 Results

In this section, the experimental results are presented to establish the contribution of each factor used to form the NPPC ensemble. First, we have pointed out the reason for using the GA for the model selection of a single NPPC. Table 2 shows the comparison between grid search and GA-based model selection method of NPPC in terms of average CPU time and testing accuracy. We have selected the 13 grid values of the four regularization parameters of NPPC as $[2^{-8}, 2^{-7}, \ldots, 2^{4}]$. Thus, to select four optimal parameters c_1, c_2, c_3, c_4, it evaluates 13^4 possible combination of parameters. It has been observed that several combinations may offer same validation accuracy. Any one of them can be used to train the final model. On the other hand for GA-based model selection, we have used the experimental setup as described in the pervious section. The results show that the GA-based model selection scheme of NPPC is very effective in terms of both computational time and classification performance. We have also reported p-values of the testing accuracy figures in 5-percent significance level, in Table 2. The p-values were calculated by performing a paired t-test [75] comparing the model selection by GA to grid search method with the assumption of the null hypothesis that there is no difference between the test set accuracy distributions of the two methods. The significant results are marked bold face in Table 2. The results show that the p-values are much less than 0.05 for all the data sets except for lung and liver cancer. This indicates that there is a significant difference in mean of the two methods. Also the time required by the GA-based model selection is approximately 10 times less than that of the grid search method. Obviously, this may vary depending upon the grid density as well as the selection of the termination criteria of the GA-based search technique. Thus, the results indicate that GA-based tuning of the model parameters of NPPC is more suitable than the tuning at discrete point by the grid search method. These results justify the use of GA for simultaneous feature and model parameter selection of NPPC.

In order to compare the performance of our NPPC ensemble, we have investigated results in terms of average accuracy and execution time required to develop an ensemble. We have compared NPPC ensemble with SVM ensemble implemented in a similar framework. The accuracy and time figures are the average of 30 independent tests with different training, testing, and validation sets. In each test, the same set of training, testing, and validation data are used for ensemble classifiers of SVM and NPPC. The single classifiers for both the cases are implemented by GA using all the genes selected by the MRMR method. Table 3 shows the improvement of the classification accuracy by ensemble classifiers than their respective single classifier for both SVM and NPPC ensembles. We have also reported the p-values of the testing accuracy figures in 5 percent significance level. Here, the p-values were calculated by performing a paired t-test comparing the ensemble method to a single classifier with the assumption of the null hypothesis that there is no difference between the test set accuracy distribution of the two methods.

From Table 3, we observe that the performances of both SVM and NPPC ensembles improve the respective single classifier. As an example, NPPC ensemble with both types of decision combination performs significantly better than the single classifier in the case of ALL_AML, lung, breast,

TABLE 3
Comparison in Terms of the Average Test Accuracy (Percent) and the Standard Deviation of Single and Ensemble Classifiers of SVM and NPPC, Respectively

Data set	No. of trained classifiers to build ensemble	Results with SVM			Results with NPPC			
		Average % test accuracy of		Av. CPU time of building SVM Ensemble (Sec.)	Average % test accuracy of			Av. CPU time of building NPPC ensemble (Sec.)
		Single SVM	SVM ensemble with majority voting / p-value		single NPPC	NPPC ensemble with majority voting/p-value	NPPC ensemble with minimum av. proximity /p-value	
ALL-AML	12	92.91±5.62	95.73±3.88 0.053677	807.36	94.52±4.65	**96.46±2.68** **0.037225**	**96.56±2.50** **0.027015**	**317.63**
Colon cancer	13	82.22±5.99	82.08±4.25 0.905221	1762.23	82.77±4.94	84.02±3.29 0.271575	83.47±2.99 0.517507	**879.65**
Lung cancer	9	97.96±0.71	**99.90±0.36** **4.05646E-8**	927.76	96.38±3.26	**99.07±1.36** **0.003326**	**99.44±0.88** **0.001389**	**292.43**
Breast cancer (ER)	8	87.87±6.01	89.82±3.37 0.097348	438.89	81.21±8.75	**91.41±4.89** **9.6260E-8**	**92.11±4.09** **3.1147E-7**	**214.86**
Lymphoma	8	81.33±12.07	**93.22±3.76** **1.6741E-5**	1453.57	86.88±6.36	**93.89±4.30** **4.8977E-6**	**94.78±2.09** **1.0235E-7**	**256.37**
Liver cancer	8	98.28±1.29	97.52±1.79 0.050850	386.34	96.77±3.71	**98.38±1.34** **0.025176**	**98.34±1.34** **0.025176**	**248.56**
Prostate cancer	7	93.33±6.46	**97.41±0.46** **0.001506**	1978.28	90.16±8.74	**97.33±0.91** **8.8624E-5**	**97.33±0.91** **8.8624E-5**	**395.49**

The figures listed in the table are the average of 30 independent runns. The p-values of the results are provided comparing the ensemble methods to a single classifier.

lymphoma, liver, and prostate cancer data sets as p-values are less than 0.05 for these cases. Only for colon cancer data set, the improvement is not significant in 5 percent significance level. On the other hand, SVM ensemble also performs significantly better than a single SVM classifier for the lung, lymphoma, and prostate cancer data sets. For other four data sets, the improvement is not statistically significant by SVM ensemble. Additionally, it is observed from Table 3 that the average time required to develop a NPPC ensemble is much less than that required to develop SVM ensemble in a similar frame work and programming environment. This is due to the computational efficiency of the linear NPPC algorithm [46].

In Table 4, we have provided the p-values of the testing accuracies by comparing the ensemble methods in the 5 percent significance level. The statistically significant results are in bold face. From the table, we observed that the overall performance of NPPC ensemble with minimum average proximity is better than both NPPC and SVM ensembles with majority voting scheme.

TABLE 4
Comparison of P-Values between the Ensemble Methods

Data set	NPPC with majority voting to SVM ensemble	NPPC with min. av. proximity to SVM ensemble	NPPC with min. av. proximity to NPPC with majority voting
ALL-AML	0.109784	**0.030079**	0.325582
Colon cancer	**1.0918E-4**	**0.022608**	**0.043397**
Lung cancer	**0.022988**	**0.019188**	**0.040569**
Breastcancer (ER)	0.118271	**0.002841**	**0.043397**
Lymphoma	0.414361	**0.036613**	0.118015
Liver cancer	0.325581	0.325581	1
Prostate cancer	**3.1085E-6**	**3.1085E-6**	1

5 Discussion

Here, we have presented a novel cancer classification method based on NPPC ensemble. The experimental results on seven data sets have demonstrated the strength of our proposed algorithm in classifying different types of cancer. Although the NPPC ensemble algorithm is assessed favorably in most cases, a drawback of the method is that there are more parameters to tune than that of SVM. But this drawback is satisfactorily removed by the GA-based tuning method.

5.1 The Effect of GA-Based Simultaneous Feature and Parameter Selection on Class Separability

In our NPPC ensemble method, we have effectively combined a wrapper method of simultaneous feature subset and model parameter selection followed by the filter method of feature selection. This offers enormous savings of computational cost of wrapper method-based feature selection. Additionally, the genes, selected by a filter method, may not be the optimal in terms of classification performance. Furthermore, for a parametric classifier, the performance is highly dependent on the selection of parameters [76]. Our NPPC is a parametric classifier and has four regularization parameters c_1, c_2, c_3, and c_4. Our goal is to assemble a number of expert classifiers in different subspaces by maximizing the 10-fold cross-validation accuracy on training data only. It is possible that the trained models may produce high cross-validation accuracy on training data. However, the question is whether the GA-based simultaneous feature and parameter selection technique can select the best combination of gene subset that retains the separability of the testing data. We

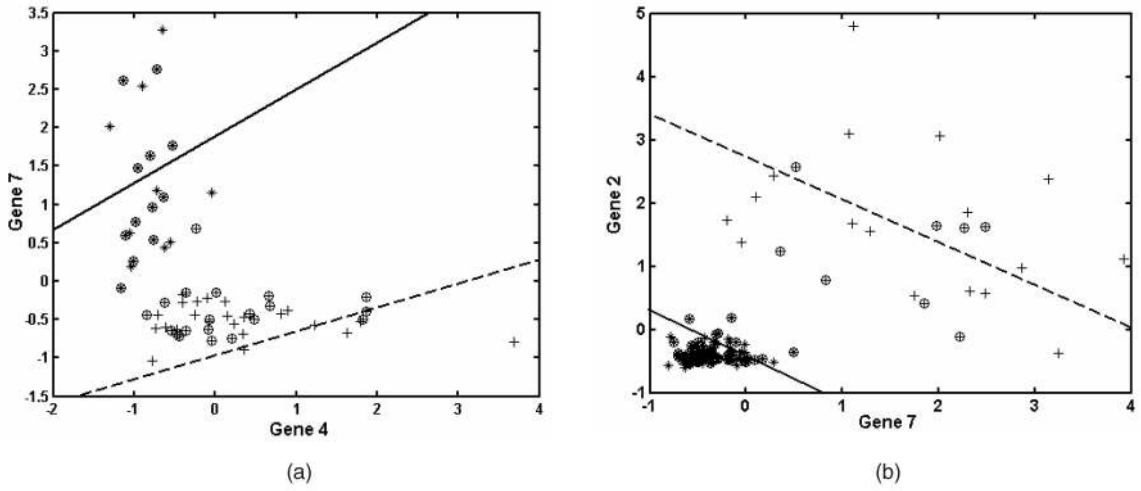

Fig. 7. The effect of the GA-based simultaneous feature and optimal parameter selection method on training and testing data. Two-dimensional NPPC expert model selected in the ensemble for (a) ALL-AML data and (b) lung cancer data. The "+" and "*" signs represent the training data of two subclasses and the circled patterns are the corresponding test data. The solid and the dotted lines represent the two hyperplanes learned from the training data.

have illustrated the effect of this method on training and testing data with two-dimensional embedding of the data points. Figs. 7a and 7b show the two-dimensional linear expert models selected in the ensemble for ALL-AML and lung cancer data sets, respectively, in experiment run 1. The two features selected for the best two-dimensional classifier are not necessarily the first two genes selected in the ranking as seen from Fig. 7. Fig. 7 also indicate that by the GA-based searching technique, the class separability of the test data may be substantially improved together with that of the training data, although the classifier building process is performed only with the training data.

5.2 The Effect of Classifier Ensemble on Classification Accuracy

The idea of combining more classifiers has been used extensively in the machine learning literature, especially for improving the classification accuracy by combining the outcome of the several classifiers. Examples are well-known bagging and boosting techniques [55, ch. 7, pp. 203-235]. However, what is innovative here is that multiple learning algorithms are not used for the purpose of improving the classification accuracy by combination or averaging but the same algorithm is used to develop best expert models in different subspaces. We have proposed a new proximity profile-based decision combination method for NPPC ensemble. The effectiveness of the proposed CME is shown in Fig. 8. The ensemble testing accuracy has been compared to the testing accuracy of each individual expert of the ensemble. We have only given the results of experiment run 1 for NPPC ensemble with the minimum average proximity method of decision combination. From Fig. 8, we observed that the individual testing accuracy of member classifiers may vary, but the ensemble accuracy is at least equal to the highest accuracy among the individual expert of the ensemble. This proves the effectiveness of the proposed combination of multiple classifiers by the minimum average proximity of the test pattern to a class.

5.3 Biological Relevance of the Selected Genes

To select a small subset of genes by the filter method, we have used the MRMR method of feature or gene selection. But our experimental design strategy evaluates several alternative training sets in order to be confident that the selected genes are statistically significant and not due to fluctuations and noise effects. We have selected those few genes which repeatedly decided on the MRMR feature ranking. To verify the biological relevance of the selected genes that are differentially expressed by this procedure, we have enlisted selected genes for the data sets in the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2010.36. Among the seven data sets, we have verified the biological relevance of the selected genes from the breast cancer data set. Out of the nine genes selected for this data set, we have found the strong biological relevance of six genes with the initiation and progression of the disease. For example, very recent research shows that the loss of expression of the transcription factor GATA3 in breast tumors has been linked to aggressive tumor development and poor patient survival [77], [78]. TFF secretory protein may have a role in bone metastasis in breast cancer [79], [80]. Genes encoding Y-box binding protein-1 (YB-1) [81], [82], [83] and C-myb [86] are oncogenes, while altered expression of oestrogen receptor (ER) [84], [87] and RSU-1/RSP-1 [85] have been shown to critically influence the cellular proliferation and cell-cycle regulation in breast cancer. Thus, our experimental procedure is able to find the genes that are relevant to the disease and can facilitate early discovery as well as prognostic and therapeutic diversification of cancer patients.

6 Conclusion

We have proposed a novel binary NPPC ensemble for gene microarray expression analysis. The new proximity profile-based CME method is found to be effective for NPPC ensemble. As a result, the NPPC ensemble technique shows good discriminating power in gene expression analysis. The NPPC ensemble provides better classification accuracy than

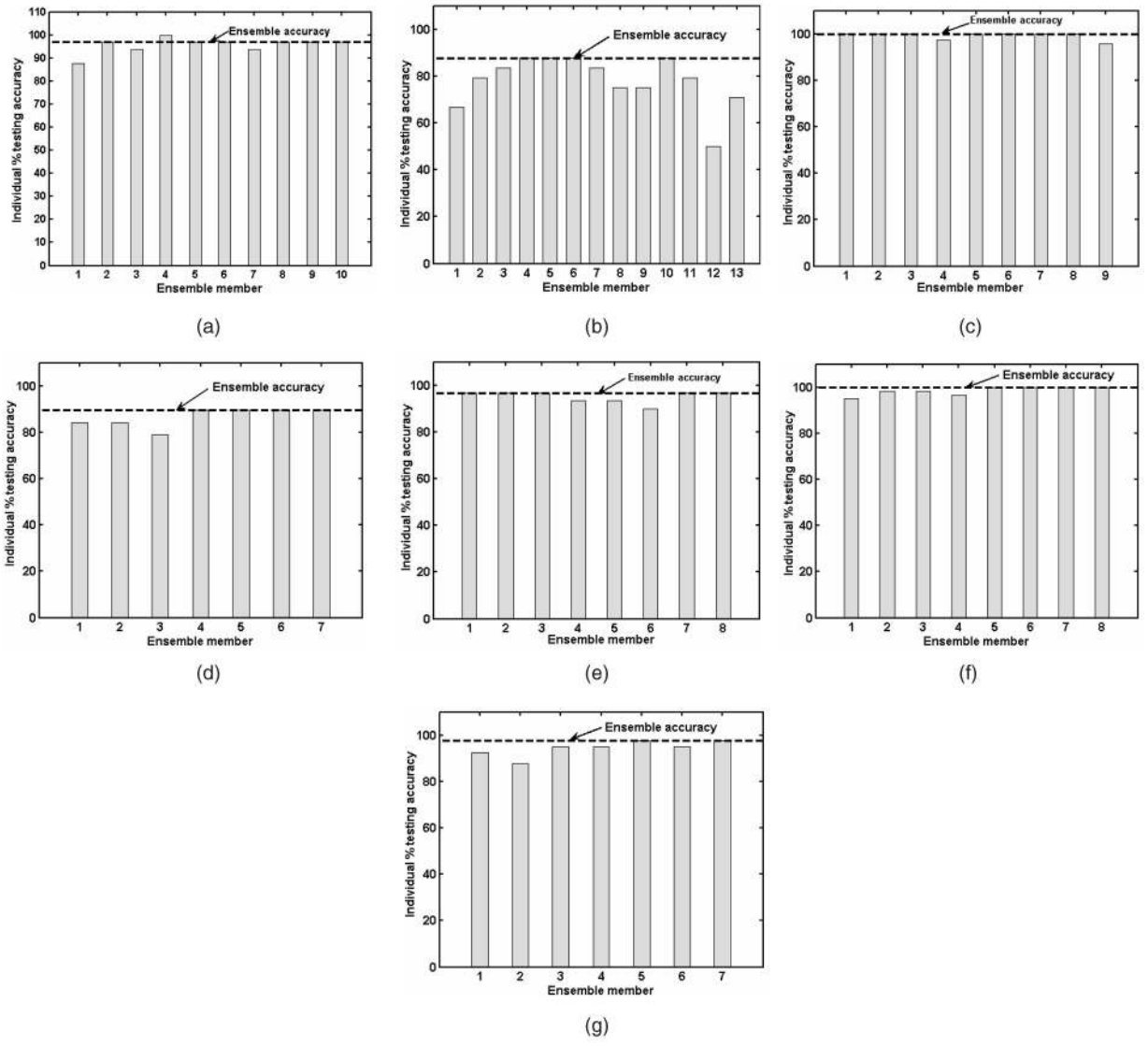

Fig. 8. Testing accuracy obtained by the NPPC ensemble with minimum average proximity and individual testing accuracy of each expert in the ensemble in the experiment of run 1 for (a) ALL_AML, (b) colon, (c) lung, (d) breast (ER), (e) lymphoma (f) liver, and (g) prostate cancer data set.

that of SVM ensemble in a similar frame work with lesser computational time. The NPPC ensemble method can be easily extended for nonlinear classifier using kernel trick. Additionally, multiclass cancer classification is possible in this frame work by extending the binary NPPC to multiclass NPPC. The proposed NPPC ensemble framework may well be of interest to others for noisy data sets in the fields of machine learning and computational biology, such as detection of horizontal gene transfer in bacterial genomes [88], classification of functional classes of proteins [89], classification of the nature of the infectious diseases [90], diagnosis of the genetic abnormalities [91], and other important areas of medical diagnostics.

APPENDIX

Appendix A is submitted as supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2010.36.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for very useful comments and suggestions which greatly improved their representation.

REFERENCES

[1] P.O. Brown and D. Botstein, "Exploring the New World of the Genome with DNA Microarrays," *Nature Genetics Supplement*, vol. 21, pp. 33-37, 1999.

[2] C. Debouck and P.N. Goodfellow, "DNA Microarrays in Drug Discovery and Development," *Nature Genetics Supplement*, vol. 21, pp. 48-50, 1999.

[3] D.J. Duggan et al., "Expression Profiling Using cDNA Micro-Arrays," *Nature Genetics Supplement*, vol. 21, pp. 10-14, 1999.

[4] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gassenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[5] B. West et al., "Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles," *Proc. Nat'l Academy of Sciences USA*, vol. 98, pp. 11462-11467, 2001.

[6] G.J. Gordon, R.V. Jenson, L.-L. Hsiao, S.R. Gullans, J.E. Blumenstock, S. Ramaswamy, W.G. Richards, D.J. Sugarbaker, and R. Bueno, "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelima," *Cancer Research,* vol. 62, pp. 4936-4967, 2002.

[7] J. Khan, J.S. Wei, and M. Ringner, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," *Nature Medicine,* vol. 7, no. 6, pp. 673-679, 2001.

[8] E. Keedwell and A. Narayanan, "Discovering Gene Networks with a Neural-Genetic Hybrid," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 2, no. 3, pp. 231-242, July-Sept. 2005.

[9] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, and W.R. Sellers, "Gene Expression Correlations of Clinical Prostate Cancer Behavior," *Cancer Cell,* vol. 1, no. 2, pp. 203-209, 2002.

[10] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning,* vol. 46, pp. 389-422, 2002.

[11] L. Li, C.R. Weinberg, T.A. Darden, and 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,* vol. 17, pp. 1131-1142, 2001.

[12] S. Kim, E.R. Dougherty, J. Barrera, Y. Chen, M. Bittner, and J.M. Trent, "Strong Feature Sets from Small Samples," *Computational Biology,* vol. 9, pp. 127-146, 2002.

[13] C. Bhattacharyya, L.R. Grate, A. Rizki, D. Radisky, F.J. Molina, M.I. Jordan, M.J. Bissell, and I.S. Mian, "Simultaneous Relevant Feature Identification and Classification in High-Dimensional Spaces: Application to Molecular Profiling Data," *Signal Processing,* vol. 83, pp. 729-743, 2003.

[14] A. Schulze and J. Downward, "Navigating Gene Expression Using Microarrays—A Technology Review," *Natural Cell Biology,* vol. 3, no. 8, pp. E190-E195, 2001.

[15] K.M. Borgwardt, S.V.N. Vishwanathan, and H. Kriegel, "Class Prediction from Time Series Gene Expression Profiles Using Dynamical Systems Kernels," *Proc. Pacific Symp. Biocomputing,* vol. 11, pp. 547-558, 2006.

[16] M. Wilson, J. DeRisi, H.H. Kristensen, P. Imboden, S. Rane, P.O. Brown, and G.K. Schoolnik, "Exploring Drug-Induced Alterations in Gene Expression in Mycobacterium Tuberculosis by Microarray Hybridization," *Proc. Nat'l Academy of Sciences USA,* vol. 96, no. 22, pp. 12833-12838, 1999.

[17] W.E. Evans and R.K. Guy, "Gene Expression as a Drug Discovery Tool," *Nature Genetics,* vol. 36, no. 3, pp. 214-215, 2004.

[18] S. Hochreiter and K. Obermayer, *Kernel Methods in Computational Biology,* B. Scholkopf, K. Tsuda, and J.-P. Vert, eds. p. 323, MIT Press, 2004.

[19] G. Bontempi, "A Blocking Strategy to Improve Gene Selection for Classification of Gene Expression Data," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 4, no. 2, pp. 293-300, Apr.-June 2007.

[20] M.A. Hall, "Correlation-Based Feature Selection Machine Learning," PhD thesis, Dept. of Computer Science, Univ. of Waikato, 1998.

[21] R. Tibshirani, T. Hastie, B. Narashiman, and G. Chu, "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proc. Nat'l Academy of Sciences USA,* vol. 99, pp. 6567-6572, 2002.

[22] J. Devore and R. Peck, *Statistics: The Exploration and Analysis of Data,* third ed. Duxbury Press, 1997.

[23] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *J. Am. Statistical Assoc.,* vol. 97, pp. 77-87, 2002.

[24] Y. Lai, B. Wu, L. Chen, and H. Zhao, "Statistical Method for Identifying Differential Gene-Gene Coexpression Patterns," *Bioinformatics,* vol. 20, pp. 3146-3155, 2004.

[25] P. Broet, A. Lewin, S. Richardson, C. Dalmasso, and H. Magdelenat, "A Mixture Model-Based Strategy for Selecting Sets of Genes in Multiclass Response Microarray Experiments," *Bioinformatics,* vol. 20, pp. 2562-2571, 2004.

[26] H. Peng, F. Long, and C. Ding, "Feature Selection on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 8, pp. 1226-1238, Aug. 2005.

[27] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence,* vol. 97, nos. 1/2, pp. 273-324, 1997.

[28] C. Huang and C. Wang, "A GA-Based Feature Selection and Parameter Optimization for Support Vector Machines," *Expert Systems with Applications,* vol. 31, pp. 231-240, 2006.

[29] Y. Tang, Y.-Q. Zhang, and Z. Huang, "Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 4, no. 3, pp. 365-381, July-Sept. 2007.

[30] M.A.T. Figuiredo and A.K. Jain, "Bayesian Learning of Sparse Classifiers," *Proc. Conf. Computer Vision and Pattern Recognition (CVPR '01),* vol. 1, pp. I-35-I-41, 2001.

[31] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning,* second ed. Springer, 2009.

[32] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays," *Statistical Science,* vol. 18, no. 1, pp. 104-117, 2003.

[33] T.S. Furey, N. Cristianini, N. Duy, D.W. Bednarski, M. Schummer, and D. Haussler, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data," *Bioinformatics,* vol. 16, no. 10, pp. 906-914, 2000.

[34] B. Krishnapuram, A.J. Hartemink, and L. Carin, "Logistic Regression and RVM for Cancer Diagnosis from Gene Expression Signatures," *Proc. IEEE Signal Processing Soc. Workshop Genomic Signal Processing and Statistics (GENSIPS),* 2002.

[35] J. Zhu and T. Hastie, "Classification of Gene Microarrays by Penalized Logistic Regression," *Biostatistics,* vol. 5, no. 2, pp. 427-443, 2004.

[36] G. Fung and O.L. Mangasarian, "Data Selection for Support Vector Machine Classifiers," *Proc. ACM SIGKDD,* pp. 64-70, 2000.

[37] L. Shen and E.C. Tan, "Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 2, no. 2, pp. 166-175, Apr.-June 2005.

[38] V. Roth, "The Generalized LASSO," *IEEE Trans. Neural Networks,* vol. 15, no. 1, pp. 16-18, Jan. 2004.

[39] V. Roth, "The Generalized LASSO: A Wrapper Approach to Gene Selection for Microarray Data," *Proc. 14th Int'l Conf. Automated Deduction (CADE),* pp. 252-255, 2002.

[40] M.A.T. Figuiredo and A.K. Jain, "Bayesian Learning of Sparse Classifiers," *Proc. Conf. Computer Vision and Pattern Recognition (CVPR '01),* 2001.

[41] B. Krishnapuram, L. Carin, and A.J. Hartemink, "Joint Classifier and Feature Optimization for Cancer Diagnosis Using Gene Expression Data," *Proc. Seventh Ann. Int'l. Conf. Computational Molecular Biology,* 2003.

[42] D. Ghosh and A.M. Chinnaiyan, "Classification and Selection of Biomarkers in Genomic Data Using Lasso," *J. Biomedical Biotechnology,* vol. 2, pp. 147-154, 2005.

[43] C. Ambroise and G. McLachlan, "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data," *Proc. Nat'l Academy of Sciences USA,* vol. 99, pp. 6562-6566, 2002.

[44] L. Wang, F. Chu, and W. Xie, "Accurate Cancer Classification Using Expressions of Very Few Genes," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 4, no. 1, pp. 40-53, Jan.-Mar. 2007.

[45] S. Ghorai, A. Mukherjee, and P.K. Dutta, "Nonparallel Plane Proximal Classifier," *Signal Processing,* vol. 89, pp. 510-522, 2009.

[46] S. Ghorai, S.J. Hossain, A. Mukherjee, and P.K. Dutta, "Newton's Method for Nonparallel Plane Proximal Classifier with Unity Norm Hyperplanes," *Signal Processing,* vol. 90, pp. 93-104, Jan. 2010.

[47] V. Vapnik, *The Nature of Statistical Learning Theory.* Springer, 1995.

[48] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines,* ch. 6, pp. 113-145. Cambridge Univ. Press, 2000.

[49] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery,* vol. 2, no. 2, pp. 121-167, 1998.

[50] O.L. Mangasarian and E.W. Wild, "Multisurface Proximal Support Vector Classification via Generalized Eigenvalues," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 1, pp. 69-74, Jan. 2006.

[51] Jayadeva, R. Khemchandani, and S. Chandra, "Twin Support Vector Machines for Pattern Classification," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 5, pp. 905-910, May 2007.

[52] V. Tresp, *Handbook for Neural Network Signal Processing,* Y.H. Hu and J.-N. Hwang, eds. CRC Press, 2001.
[53] V. Tresp, "A Bayesian Committee Machine," *Neural Computation,* vol. 12, pp. 2719-2741, 2000.
[54] D. Martinez and G. Millerioux, "Support Vector Committee Machines," *Proc. European Symp. Artificial Neural Networks (ESANN '00),* pp. 43-48, 2000.
[55] L.I. Kuncheva, *Combining Pattern Classifiers—Methods and Algorithms.* Wiley Interscience, 2004.
[56] M. Aksela and J. Laaksonen, "Using Diversity of Errors for Selecting Members of a Committee Classifier," *Pattern Recognition,* vol. 39, pp. 608-623, 2006.
[57] J. Yao, R.M. Summers, and A. Hara, "Optimizing the Support Vector Machines (SVM) Committee Configuration in a Colonic Polyp CAD System," *Proc. SPIE Conf.,* 2005.
[58] Y.-W. Kim and I.-S. Oh, "Classifier Ensemble Selection Using Hybrid Genetic Algorithms," *Pattern Recognition Letters,* vol. 29, pp. 796-802, 2008.
[59] G. Rogova, "Combining the Results of Several Neural Network Classifiers," *Neural Networks,* vol. 7, pp. 777-781, 1994.
[60] Y.S. Huang and C.Y. Suen, "A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 1, pp. 90-94, Jan. 1995.
[61] R.L. Haupt and S.E. Haupt, *Practical Genetic Algorithms,* second ed. Wiley Interscience, 2004.
[62] J. Nocedal and S. Wright, *Numerical Optimization,* second ed. Springer, 2006.
[63] D.P. Bertsekas, *Nonlinear Programming,* second ed. Athena Scientific, 1999.
[64] Y.-J. Lee and O.L. Mangasarian, "SSVM: A Smooth Support Vector Machine," *Computational Optimization and Application,* vol. 20, pp. 5-22, 2001.
[65] R. Kohavi, "A Study of Cross-Validation and Boot Strap for Accuracy Estimation and Model Selection," *Proc. Int'l Joint Conf. Artificial Intelligence,* 1995.
[66] M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn, and A.K. Jain, "Dimensionality Reduction Using Genetic Algorithms," *IEEE Trans. Evolutionary Computation,* vol. 4, no. 2, pp. 164-171, July 2000.
[67] W.F. Punch, E.D. Goodman, M. Pei, L. Chia-Shun, P. Hovland, and R. Enbody, "Further Research on Feature Selection and Classification Using Genetic Algorithms," *Proc. Int'l Conf. Genetic Algorithms,* pp. 557-564, 1993.
[68] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissue Probed by Oligonucleotide Arrays," *Proc. Nat'l Academy of Sciences USA,* vol. 96, pp. 6745-6750, 1999.
[69] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C.T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, and T.R. Golub, "Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene Expression Profiling and Supervised Machine Learning," *Nature Medicine,* vol. 8, pp. 68-74, 2002.
[70] X. Chen et al., "Gene Expression Patterns in Human Liver Cancers," *Molecular Biology of the Cell,* vol. 13, pp. 1929-1939, 2002.
[71] H. Xiong, Y. Zhang, and X.-W. Chen, "Data-Dependent Kernel Machines for Microarray Data Classification," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 4, no. 4, pp. 583-595, Oct.-Dec. 2007.
[72] O. Troyanskaya et al., "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics,* vol. 17, pp. 520-525, 2001.
[73] *MATLAB, User's Guide,* The MathWorks Inc., http://www.mathworks.com, 1994-2009.
[74] S.R. Gunn, Support Vector Machine Matlab Toolbox, http://www.isis.ecs.soton.ac.uk/resources/svminfo/, 1998.
[75] T.M. Mitchell, *Machine Learning,* ch. 5, p. 148. McGraw-Hill, 1997.
[76] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification.* second ed. John Wiley & Sons, 2001.
[77] A.B. Dydensborg, A.A. Rose, B.J. Wilson, D. Grote, M. Paquet, V. Giguère, P.M. Siegel, and M. Bouchard, "GATA3 Inhibits Breast Cancer Growth and Pulmonary Breast Cancer Metastasis," *Oncogene,* vol. 28, pp. 2634-2642, July 2009.
[78] J.M. Arnold, D.Y.H. Choong, E.R. Thompson, k. ConFab, N. Waddell, G.J. Lindeman, J.E. Visvader, I.G. Campbell, and G.C. Trench, "Frequent Somatic Mutations of GATA3 in Non-BRCA1/BRCA2 Familial Breast Tumors, but Not in BRCA1-, BRCA2- or Sporadic Breast Tumors," *Breast Cancer Research and Treatment,* vol. 119, no. 2, pp. 491-496, Feb. 2010.
[79] M. Smid, Y. Wang, J.G.M. Klijn, A.M. Sieuwerts, Y. Zhang, D. Atkins, J.W.M. Martens, and J.A. Foekens, "Genes Associated with Breast Cancer Metastatic to Bone," *J. Clinical Oncology,* vol. 24, no. 15, pp. 2261-2267, 2006.
[80] K. Tjensvoll, B. Gilje, S. Oltedal, V.F. Shammas, J.T. Kvaly, R. Heikkilä1, and O. Nordgård, "A Small Subgroup of Operable Breast Cancer Patients with Poor Prognosis Identified by Quantitative Real-Time RT-PCR Detection of Mammaglobin A and Trefoil Factor 1 mRNA Expression in Bone Marrow," *Breast Cancer Research and Treatment,* vol. 116, no. 2, pp. 329-338, 2009.
[81] A. Astanehe, M.R. Finkbeiner, P. Hojabrpour, K. To, A. Fotovati, A. Shadeo, A.L. Stratford, W.L. Lam, I.M. Berquin, V. Duronio, and S.E. Dunn, "The Transcriptional Induction of PIK3CA in Tumor Cells Is Dependent on the Oncoprotein Y-Box Binding Protein-1," *Oncogene,* vol. 28, pp. 2406-2418, June 2009.
[82] G. Habibi, S. Leung, J.H. Law, K. Gelmon, H. Masoudi, D. Turbin, M. Pollak, T.O. Nielsen, D. Huntsman, and S.E. Dunn, "Redefining Prognostic Factors for Breast Cancer: YB-1 Is a Stronger Predictor of Relapse and Disease-Specific Survival than Estrogen Receptor or HER-2 across All Tumor Subtypes," *Breast Cancer Research,* vol. 10, no. 5, Oct. 2008.
[83] T. Fujii, A. Kawahara, Y. Basaki, S. Hattori, K. Nakashima, K. Nakano, K. Shirouzu, K. Kohno, T. Yanagawa, H. Yamana, K. Nishio, M. Ono, M. Kuwano, and M. Kage, "Expression of Her-2 and Estrogen Receptor Alpha Depends upon Nuclear Localization of Y-Box Binding Protein-1 in Human Breast Cancers," *Cancer Research,* vol. 68, no. 5, pp. 1504-1512, 2008.
[84] S. Heck, J. Rom, V. Thewes, N. Becker, B. Blume, H.P. Sinn, U. Deuschle, C. Sohn, A. Schneeweiss, and P. Lichter, "Estrogen-Related Receptor {Alpha} Expression and Function Is Associated with the Transcriptional Coregulator AIB1 in Breast Carcinoma," *Cancer Research,* vol. 69, no. 12, pp. 5186-5193, 2009.
[85] F. Vasaturo, G.W. Dougherty, and M.L. Cutler, "Ectopic Expression of Rsu-1 Results in Elevation of p21CIP and Inhibits Anchorage-Independent Growth of MCF7 Breast Cancer Cells," *Breast Cancer Research and Treatment,* vol. 61, no. 1, pp. 69-78, 2000.
[86] F. Fang, M.A. Rycyzyn, and C.V. Clevenger, "Role of c-Myb during Prolactin-Induced Signal Transducer and Activator of Transcription 5a Signaling in Breast Cancer Cells," *Endocrinology,* vol. 150, no. 4, pp. 1597-1606, 2009.
[87] R.G. Ramsay and T.J. Gonda, "MYB Function in Normal and Cancer Cells," *Nature Rev. Cancer,* vol. 8, no. 7, pp. 523-524, 2008.
[88] H. Ning, B. Yang, J. Cui, and L. Jing, "Detection of Horizontal Gene Transfer in Bacterial Genomes," *Proc. Third Int'l Symp. Optimization and System Biology (OSB '09),* pp. 229-236, Sept. 2009.
[89] L. Chen, L. Lu, K. Feng, W. Li, J. Song, L. Zheng, Y. Yuan, Z. Zeng, K. Feng, W. Lu, and Y. Cai, "Multiple Classifier Integration for the Prediction of Protein Structural Classes," *J. Computational Chemistry,* vol. 30, no. 14, pp. 2248-2254, 2009.
[90] M.T. Cordeiro, U. Barga-Neto, R.M. Noqueira, and E.T. Marques, "Reliable Classifier to Differentiate Primary and Secondary Acute Dengue Infection Based on IgG ELISA," *Public Library of Science (PLoS) One,* vol. 4, no. 4, Apr. 2009.
[91] B. Lerner, J. Yeshaya, and L. Koushnir, "On the Classification of a Small ~Imbalanced Cytogenetic Image Database," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 4, no. 2, pp. 204-215, Apr.-June 2007.

Santanu Ghorai received the BSc degree in physics (Hons.) and the BTech degree in instrumentation engineering from Calcutta University, Kolkata, in 1995 and 1998, respectively. He received the ME degree in electrical engineering from Jadavpur University, Kolkata, in 2000. From 2000-2006, he was associated as a faculty member at the MCKV Institute of Engineering, Howrah, West Bengal, India, in the Department of Electronics and Communication Engineering. From July 2006 to July 2009, he was working toward his PhD degree as a QIP research scholar in the Department of Electrical Engineering, Indian Institute of Technology, Kharagpur, India. He received his PhD degree in Nov. 2010. He rejoined the Department of Electronics and Communication Engineering at the MCKV Institute of Engineering in July 2009. His main research interests include pattern classification in computational biology. He is a student member of the IEEE.

Anirban Mukherjee received the BEE degree in electrical engineering from Jadavpur University, Kolkata, in 1998. He received the MTech and PhD degrees in electrical engineering from the Indian Institute of Technology, Kharagpur. He was with the Centre of Excellence for Embedded Systems, Tata Consultancy Services, in 2004-2005. He is currently with the faculty of Electrical Engineering Department at the Indian Institute of Technology, Kharagpur. His main research interests include image processing and pattern classification in computational biology. He is a member of the IEEE.

Sanghamitra Sengupta received the MSc and PhD degrees in 1991 and 1998, respectively, from the University of Calcutta. After obtaining postdoctoral training in the Department of Genetics of Case Western Reserve University and Stanford University, she joined the Human Genetics Unit of the Indian Statistical Institute as a research scientist. She has been a faculty member in the Department of Biochemistry, University of Calcutta, since 2005. She is also a member of the Indian Society of Human Genetics and the Society of Biological chemists. Her main research interests include population genetics and cancer genomics. Her work has resulted in 10 research papers in the field of genetics.

Pranab K. Dutta received the BE, ME, and PhD degrees, all in electrical engineering, in 1984, 1986, and 1992, respectively. Since 1993, he has been a faculty member in the Department of Electrical Engineering, Indian Institute of Technology, Kharagpur. He is currently the head of the School of Medical Science and Technology, Indian Institute of Technology, Kharagpur. His main interests include signal and image processing and biomedical instrumentation. He has published more than 100 research papers. His work in these areas resulted in more than eight patents and patent applications.