Relevant and Significant Supervised Gene Clusters for Microarray Cancer Classification

Pradipta Maji* and Chandra Das

Abstract—**An important application of microarray data in functional genomics is to classify samples according to their gene expression profiles such as to classify cancer versus normal samples or to classify different types or subtypes of cancer. One of the major tasks with gene expression data is to find co-regulated gene groups whose collective expression is strongly associated with sample categories. In this regard, a gene clustering algorithm is proposed to group genes from microarray data. It directly incorporates the information of sample categories in the grouping process for finding groups of co-regulated genes with strong association to the sample categories, yielding a supervised gene clustering algorithm. The average expression of the genes from each cluster acts as its representative. Some significant representatives are taken to form the reduced feature set to build the classifiers for cancer classification. The mutual information is used to compute both gene-gene redundancy and gene-class relevance. The performance of the proposed method, along with a comparison with existing methods, is studied on six cancer microarray data sets using the predictive accuracy of naive Bayes classifier, K-nearest neighbor rule, and support vector machine. An important finding is that the proposed algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability.**

Index Terms—**Classification, feature selection, gene clustering, microarray analysis, mutual information.**

I. Introduction

RECENT advancement of microarray technology has made the experimental study of gene expression data faster and more efficient. Microarray techniques such as DNA and high density oligonucleotide chips are powerful biotechnologies as they are able to record the expression levels of thousands of genes simultaneously. The vast amount of gene expression data leads to statistical and analytical challenges including the classification of the data set into correct classes. Hence, one of the important applications of gene expression data in functional genomics is to classify samples according to their gene expression profiles such as to classify cancer versus normal samples or to classify different types of cancer [1].

A microarray gene expression data set can be represented by an expression table, where each row corresponds to one particular gene, each column to a sample, and each entry of the matrix is the measured expression level of a particular gene in a sample, respectively [1], [2]. However, the major problem of microarray gene expression data-based sample classification is the huge number of genes compared to the limited number of samples. Most classification algorithms suffer from such a high dimensional input space. Furthermore, most of the genes in arrays are irrelevant to sample classification. These genes may also introduce noises and decrease prediction accuracy. In addition, a biomedical concern for researchers is to identify the key marker genes, which discriminate samples for class diagnosis. Therefore, the gene selection is crucial for sample classification in medical diagnostics as well as for understanding how the genome as a whole works [3], [4]. As this is a feature selection problem, clustering method can be used that partitions the given gene set into subgroups, each of which should be as homogeneous as possible [5]–[7].

The conventional clustering methods such as hierarchical clustering [8], k-means algorithm [9], and self organizing map [10] group a subset of genes that are interdependent with each other. In other words, genes in a cluster are more correlated with each other, whereas genes in different clusters are less correlated [7], [11]. The gene clustering is able to reduce search dimension of a classification algorithm and constructs the model using a tightly correlated subset of genes rather than using the entire gene space. After clustering genes, a reduced set of genes can be selected for further analysis. However, these algorithms usually fail to reveal functional groups of genes that are of special interest in sample classification as the genes are clustered by similarity only, without using any information about the sample categories [12].

To reveal groups of co-regulated genes with strong association to the sample categories, different supervised gene clustering algorithms have been proposed recently [12]–[15]. The supervised gene clustering is defined as the grouping of genes, controlled by the values of genes as well as the supervised information of sample categories [12], [15]. Previous work in this field encompasses tree harvesting [13], a two step method which consists first of generating numerous candidate groups by unsupervised hierarchical clustering. Then, the average expression profile of each cluster is considered as a potential input variable for a response model and the few gene groups that contain the most useful information for tissue discrimination are identified. Only this second step makes the clustering supervised, as the selection process relies on external information about tissue types.

An interesting supervised clustering approach that directly incorporates the response variables in the grouping process is the partial least squares procedure [14], which in a supervised

Manuscript received August 03, 2011; revised March 22, 2012; accepted March 27, 2012. Date of publication April 27, 2012; date of current version May 30, 2012. *Asterisk indicates corresponding author.*

*P. Maji is with the Machine Intelligence Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata, 700 108, India (e-mail: pmaji@isical.ac.in).

C. Das is with the Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, 700 152, India.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

manner constructs weighted linear combinations of genes that have maximal covariance with the outcome. However, it has the drawback that the fitted components involve all, usually thousands of genes, which makes them very difficult to interpret. Moreover, partial least squares for every component yields a linear combination of gene expressions which completely lacks the biological interpretation of having a cluster of genes acting similarly in the same pathway.

A direct approach to combine gene selection, clustering and supervision in one single step is reported in [12]. The supervised gene clustering algorithm proposed in [12] is a combination of gene selection for cluster membership and formation of a new predictor by possible sign flipping and averaging the gene expressions within a cluster. The cluster membership is determined with a forward and backward searching technique that optimizes the Wilcoxon test based predictive score and margin criteria defined in [12], which both involve the supervised response variables from the data. However, as both predictive score and margin criteria depend on the actual gene expression values, they are very much sensitive to noise or outlier of the data set.

In this regard, a new supervised gene clustering algorithm is proposed to find co-regulated clusters of genes whose collective expression is strongly associated with the sample categories or class labels. The mutual information is used to measure both gene-gene similarity and gene-class relevance. The proposed algorithm uses this measure to reduce the redundancy among genes. It involves partitioning of the original gene set into some distinct subsets or clusters so that the genes within a cluster are highly co-regulated with strong association to the sample categories while those in different clusters are as dissimilar as possible. A single gene from each cluster having the highest gene-class relevance value is first selected as the initial representative of that cluster. The representative of each cluster is then modified by averaging the initial representative with other genes of that cluster whose collective expression is strongly associated with the sample categories. It is based on sequentially improving the gene-class relevance value of the cluster representative that measures the clusters' strength for discriminating the sample categories. In effect, the proposed algorithm yields biologically significant gene clusters whose coherent average expression levels allow perfect discrimination of sample categories. After generating all clusters and their representatives, a few representatives are selected based on their class discrimination power and are passed through the classifier to classify samples. The performance of the proposed method, along with a comparison with exiting methods, is studied on six cancer microarray data sets using the predictive accuracy of support vector machine (SVM), K-nearest neighbor (K-NN) rule, and naive Bayes (NB) classifier. The results demonstrate that the proposed method is more effective for microarray cancer classification.

II. Feature Extraction for Cancer Classification

This section presents a gene clustering algorithm for identifying relevant and significant supervised gene clusters from microarray gene expression data to classify cancer samples.

A. Proposed Gene Clustering Method

The basic stochastic model of the proposed algorithm for microarray data equipped with categorical response is given by a random pair

$$(\xi, \mathbb{D}) \text{ with values } \Re^m \times \mathbb{D} \tag{1}$$

where $\xi \in \Re^m$ denotes a log-transformed gene expression profile of a tissue sample standardized to zero mean and unit variance and $\mathbb{D} = \{1, \cdots, i, \cdots, \mathrm{K}\}$ is the associated response variable, where K represents the number of classes.

Let $\mathbb{G} = \{\mathcal{X}_1, \cdots, \mathcal{X}_i, \cdots, \mathcal{X}_m\}$ represents the full gene set of a given microarray data. To account for the fact that not all m genes on the chip, but rather a few functional gene subsets, determine nearly all of the outcome variation and thus the type of a tissue, the whole gene set is partitioned into k number of functional groups or clusters $\mathbb{C} = \{\mathbb{C}_1, \cdots, \mathbb{C}_k\}$ with $k \ll m$. They form usually an incomplete partition of the gene set: $\{\cup_{i=1}^{k} \mathbb{C}_i\} \subset \{1, \cdots, m\}$ and $\mathbb{C}_i \cap \mathbb{C}_j \neq \emptyset, i \neq j$. Finally, a representative of each cluster is generated and among them a few forms the reduced feature set. Let $\bar{\mathcal{X}}_i \in \Re$ denotes a representative expression value of gene cluster $\mathbb{C}_i$. There are many possibilities to determine such group values $\bar{\mathcal{X}}_i$, but as one would like to shape clusters that contain similar genes, a simple linear combination is an accurate choice:

$$\bar{\mathcal{X}}_i = \frac{1}{|\mathbb{C}_i|} \sum_{\mathcal{X}_j \in \mathbb{C}_i} \epsilon_j \mathcal{X}_j, \qquad \epsilon_j \in \{-1, 1\}. \tag{2}$$

Because of the use of log-transformed, mean-centered, and standardized expression data, the contribution of a particular gene $\mathcal{X}_j$ to the group value $\bar{\mathcal{X}}_i$ is also allowed to be given by its sign-flipped expression values $-\mathcal{X}_j$. This means that both under and over expressions are treated symmetrically, and it prevents the differential expression of genes with different polarity from canceling out when they are tagged.

The proposed clustering algorithm relies on mainly two factors, namely, determining the relevance of each gene and growing the cluster around each relevant gene incrementally by adding one gene after the other. One of the important property of the proposed clustering approach is that the cluster is augmented by the genes those satisfy following two conditions, namely, suit best into the current cluster in terms of similarity measure; and improve the differential expression of the current cluster most, according to the relevance of the cluster representative or prototype. The growth of a cluster is repeated until the cluster stabilizes, and then the proposed algorithm starts to generate a new cluster.

Let $\mathrm{R}_{\mathcal{X}_i}(\mathbb{D})$ represents the relevance of gene $\mathcal{X}_i \in \mathbb{G}$ with respect to the class label $\mathbb{D}$. The relevance uses information about the class labels and is thus a criterion for supervised clustering. The proposed algorithm starts with a single gene $\mathcal{X}_i$ of $\mathbb{G}$ that has the highest relevance value with respect to the class labels. The cluster $\mathbb{C}_i$ is then formed by selecting the set of genes $\{\mathcal{X}_j\}$ from the whole set $\mathbb{G}$ considering the gene $\mathcal{X}_i$ as the initial representative of cluster $\mathbb{C}_i$, where each gene $\mathcal{X}_j \in \mathbb{C}_i$ must satisfy following two conditions:

1) the similarity between the initial cluster representative $\mathcal{X}_i$ and gene $\mathcal{X}_j$, $\psi(\mathcal{X}_i, \mathcal{X}_j) \geq \delta$, where the threshold δ is termed as the radius of cluster $\mathbb{C}_i$; and
2) merging of gene $\mathcal{X}_j$ with the current cluster representative increases the relevance value.

The representative of cluster $\mathbb{C}_i$ is refined incrementally. By searching among the genes of $\mathbb{G}$, the current cluster representative is merged and averaged with one single gene whose similarity value with initial representative $\mathcal{X}_i$ is greater than δ and that increases the relevance value of the augmented cluster representative $\bar{\mathcal{X}}_i$. The merging process is repeated until the relevance value can no longer be improved. Instead of considering all genes of $\mathbb{G}$, the augmented gene $\bar{\mathcal{X}}_i$ is computed by considering a subset of genes whose similarity values with initial representative $\mathcal{X}_i$ is greater than δ. The set of genes $\mathbb{C}_i$ satisfying above two conditions represents the cluster associated with the gene $\mathcal{X}_i$.

The generation of cluster reduces the redundancy among genes of the set $\mathbb{G}$ as well as increases the relevance with respect to class labels. After generating the augmented cluster representative $\bar{\mathcal{X}}_i$ from the cluster $\mathbb{C}_i$, the process is repeated to find more clusters and augmented cluster representatives by discarding the gene $\mathcal{X}_i$ and all other genes $\mathcal{X}_j \in \mathbb{C}_i$ from the set $\mathbb{G}$ whose similarity values with $\mathcal{X}_i$, $\psi(\mathcal{X}_i, \mathcal{X}_j) \geq \beta$. After generating all k clusters and their representatives, the best d cluster representatives are selected according to their class relevance value and are passed through classifiers to generate classification rule. The main steps of the proposed gene clustering algorithm are reported in Fig. 1.

The main advantage of the proposed algorithm over conventional methods is that it generates clusters based on gene expression values as well as supervised information of sample categories. In effect, it can find functional groups of genes that are of most important in sample classification. Unlike the supervised gene clustering algorithm of [12], the proposed algorithm considers only a subset of genes, instead of whole gene set, to form a cluster whose similarity values are greater than a given threshold. Also, the proposed method generates overlapping supervised gene clusters for sample classification, unlike the method of Maji [15], based on the fact that many genes may participate in multiple different processes as the biological processes are not independent of each other. In effect, it is able to extract features that contain sound biological information and are also very powerful class discriminator.

In this work, mutual information is used to measure the gene-class relevance and gene-gene similarity. In principle, mutual information is used to quantify the information shared by two objects. If two independent objects do not share much information, mutual information value between them is small. While two highly correlated objects will demonstrate a high mutual information value [16]. The objects can be the class label and genes. The necessity for a gene to be an independent and informative can be determined by the shared information between the gene and rest as well as that between the gene and class label. If a gene has expression values randomly distributed in different classes, its mutual information with these classes is zero. If a gene is strongly differentially expressed for different classes, it should have large mutual information. Hence, mutual information can be used as the measures of both gene-class relevance and gene-gene similarity or redundancy [4], [15], [17], [18].

In microarray gene expression data sets, the class labels of samples are represented by discrete symbols, while the expression values of genes are continuous. Hence, to measure both gene-class relevance of a gene with respect to class labels and gene-gene redundancy between two genes using mutual information, the continuous expression values of a gene are usually divided into several discrete partitions. The a prior or marginal probabilities and their joint probabilities are then calculated to compute both gene-class relevance and gene-gene redundancy using the definitions for discrete cases. In this paper, the discretization method reported in [4], [17] is employed to discretize continuous expression values.

Input:	An $m \times n$ gene expression matrix $\mathcal{T} = \{\xi_{ij}\}_{m \times n}$. m and n represent number of genes and samples.
Output:	A set containing d augmented cluster representatives.
Definitions:	
$\mathbb{G}$	set of genes $\{\mathcal{X}_1, \cdots, \mathcal{X}_m\}$ of microarray data
$\mathrm{R}_{\mathcal{X}_i}(\mathbb{D})$	relevance value of gene $\mathcal{X}_i$ with respect to class $\mathbb{D}$
$\psi(\mathcal{X}_i, \mathcal{X}_j)$	similarity value between genes $\mathcal{X}_i$ and $\mathcal{X}_j$
$\mathbb{C}_i$	cluster associated with the gene $\mathcal{X}_i$
$\bar{\mathcal{X}}_i$	augmented gene of cluster $\mathbb{C}_i$
k	number of clusters to be generated
$\mathbb{S}$	set of augmented genes $\{\bar{\mathcal{X}}_1, \cdots, \bar{\mathcal{X}}_k\}$
δ, β	user defined thresholds
d	number of representatives to be selected for classifier

1) initialize $\mathbb{G}$ to $\{\mathcal{X}_1, \cdots, \mathcal{X}_m\}$, $\mathbb{S}$ to $\emptyset$.
2) do for $i = 1$ to m
   - compute relevance value $\mathrm{R}_{\mathcal{X}_i}(\mathbb{D})$ of gene $\mathcal{X}_i \in \mathbb{G}$.
3) do until $\mathbb{G} = \emptyset$ or k gene clusters are generated
   a) set $\mathrm{R}_{\max}$ and i to zero.
   b) do for $j = 1$ to $|\mathbb{G}|$
      - if $\mathrm{R}_{\mathcal{X}_j}(\mathbb{D}) \geq \mathrm{R}_{\max}$, then set $\mathrm{R}_{\max}$ to $\mathrm{R}_{\mathcal{X}_j}(\mathbb{D})$ and i to j.
   c) add $\mathcal{X}_i$ to $\mathbb{C}_i$ and delete $\mathcal{X}_i$ from $\mathbb{G}$.
   d) initialize $\bar{\mathcal{X}}_i$ to $\mathcal{X}_i$.
   e) do for $j = 1$ to $|\mathbb{G}|$
      e.1) compute similarity $\psi(\mathcal{X}_i, \mathcal{X}_j)$ between $\mathcal{X}_i$ and $\mathcal{X}_j$.
      e.2) if $\psi(\mathcal{X}_i, \mathcal{X}_j) \geq \delta$, then
         i) compute two augmented representatives $\bar{\mathcal{X}}^{+}_{i+j}$ and $\bar{\mathcal{X}}^{-}_{i+j}$ by averaging $\mathcal{X}_j$ and its complement with genes of $\mathbb{C}_i$.
         ii) compute relevance $\mathrm{R}_{\bar{\mathcal{X}}^{+}_{i+j}}(\mathbb{D})$ and $\mathrm{R}_{\bar{\mathcal{X}}^{-}_{i+j}}(\mathbb{D})$.
         iii) if $\mathrm{R}_{\bar{\mathcal{X}}^{+}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathcal{X}}^{-}_{i+j}}(\mathbb{D})$, then set $\bar{\mathcal{X}}_{i+j}$ to $\bar{\mathcal{X}}^{+}_{i+j}$ and $\mathrm{R}_{\bar{\mathcal{X}}_{i+j}}(\mathbb{D})$ to $\mathrm{R}_{\bar{\mathcal{X}}^{+}_{i+j}}(\mathbb{D})$ else set $\bar{\mathcal{X}}_{i+j}$ to $\bar{\mathcal{X}}^{-}_{i+j}$ and $\mathrm{R}_{\bar{\mathcal{X}}_{i+j}}(\mathbb{D})$ to $\mathrm{R}_{\bar{\mathcal{X}}^{-}_{i+j}}(\mathbb{D})$.
         iv) if $\mathrm{R}_{\bar{\mathcal{X}}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathcal{X}}_i}(\mathbb{D})$, then set $\bar{\mathcal{X}}_i$ to $\bar{\mathcal{X}}_{i+j}$, $\mathrm{R}_{\bar{\mathcal{X}}_i}(\mathbb{D})$ to $\mathrm{R}_{\bar{\mathcal{X}}_{i+j}}(\mathbb{D})$, add $\mathcal{X}_j$ to $\mathbb{C}_i$.
         v) if $\psi(\mathcal{X}_i, \mathcal{X}_j) \geq \beta$, then delete $\mathcal{X}_j$ from $\mathbb{G}$.
   f) add $\bar{\mathcal{X}}_i$ to $\mathbb{S}$.
4) do until $\mathbb{S} = \emptyset$ or d representatives are selected.
   a) set $\mathrm{R}_{\max}$ and i to zero.
   b) do for $j = 1$ to $|\mathbb{S}|$
      - if $\mathrm{R}_{\bar{\mathcal{X}}_j}(\mathbb{D}) \geq \mathrm{R}_{\max}$, then set $\mathrm{R}_{\max}$ to $\mathrm{R}_{\bar{\mathcal{X}}_j}(\mathbb{D})$ and i to j.
   c) select $\bar{\mathcal{X}}_i$ from $\mathbb{S}$ and delete $\bar{\mathcal{X}}_i$ from $\mathbb{S}$.
5) end.

Fig. 1. Main steps of proposed gene clustering algorithm.

B. Computational Complexity

The computation of the relevance of m genes is carried out in step 2 of the proposed algorithm, which has $\mathcal{O}(mn)$ time complexity as the time required to compute the relevance of each gene is n; n being the number of samples. The cluster generation step, that is step 3, is executed k times to generate k clusters and corresponding augmented cluster representatives. There are two loops in the cluster generation step; each of them is executed m times. Each iteration of the first loop, that is step 3(b), takes only a constant amount of time. Hence, the complexity of this step is $\mathcal{O}(m)$. On the other hand, three major tasks, namely, computation of similarity between two genes, that of two augmented representatives, and their relevance values, are performed within the second loop, that is step 3(e), which have overall $\mathcal{O}(n)$ time complexity. Other tasks take only a constant amount of time. Hence, the complexity to generate k clusters using step 3 is $\mathcal{O}(k(m+mn))$, that is, $\mathcal{O}(kmn)$. Finally, step 4 performs the selection of d augmented cluster representatives according to their relevance values from the k augmented representatives, which has a computational complexity of $\mathcal{O}(kd)$. Hence, the overall time complexity of the proposed supervised gene clustering algorithm is $\mathcal{O}(mn+kmn+kd)$, that is, $\mathcal{O}(k(mn+d))$. However, as $k, d, n \ll m$, the proposed clustering algorithm has an overall $\mathcal{O}(m)$ time complexity.

III. Experimental Results and Discussion

The performance of the proposed mutual information based supervised gene clustering (MSG) algorithm is extensively compared with that of some existing supervised and unsupervised gene clustering and gene selection algorithms, namely, attribute clustering algorithm (ACA) [7], supervised gene clustering algorithm (SGCA) [12], and minimum redundancy-maximum relevance (mRMR) framework [17]. To analyze the performance of different algorithms, the experimentation is done on six cancer microarray gene expression data sets. The major metric for evaluating the performance of different algorithms is the classification accuracy of the NB classifier [6], K-NN rule [6], and SVM [19].

To compute the classification accuracy, both leave-one-out cross-validation (LOOCV) and bootstrap approach are performed on each gene expression data set. For each training set, a set of gene clusters and their augmented representatives are first generated, and then one of the classifiers is trained with the augmented representatives. After the training, the features for the test sample are first constructed using the information of genes those were used to generate augmented representatives for the training set and then the class label of the test sample is predicted using the classifier. In all experiments, maximum fifty clusters ($k = 50$) and their corresponding representatives are generated. Among them three best representatives ($d = 3$) are selected for analysis.

A. Gene Expression Data Sets

In this paper, publicly available following six cancer data are used since binary classification is a typical and fundamental issue in diagnostic and prognostic prediction of cancer.

1) Breast Cancer I: The breast cancer data set contains expression levels of 7129 genes in 49 breast tumor samples [20]. The samples are classified according to their estrogen receptor (ER) status: 25 samples are ER positive while the other 24 samples are ER negative.

2) Leukemia: It is an affymetrix high-density oligonucleotide array that contains 7070 genes and 72 samples from two classes of leukemia [1]: 47 acute lymphoblastic leukemia and 25 acute myeloid leukemia.

3) Colon Cancer: The colon cancer data set contains expression levels of 2000 genes and 62 samples from two classes [21]: 40 tumor and 22 normal colon tissues.

4) Lung Cancer: This data set contains 181 tissue samples: among them 31 are malignant pleural mesothelioma and rest 150 adenocarcinoma of the lung [22]. Each sample is described by the expression levels of 12533 genes.

5) Breast Cancer II: In this data set, relapse or non relapse of metastases in patients after initial diagnosis for interval of at least 5 years has been classified in breast cancer patients [23]. Total 97 samples are given: 46 patients developed distance metastases within 5 years, labeled as relapse, while 51 remained healthy, labeled as non-relapse. The data set consists of 24 188 genes.

6) Prostate Cancer: In this data set, 136 samples are grouped into two classes: 77 prostate tumor and 59 prostate normal samples [24]. Each sample contains 12 600 genes.

B. Class Prediction Methods

Following three classifiers are used to evaluate the performance of the proposed clustering algorithm.

1) SVM: The SVM [19] is a margin classifier that draws an optimal hyperplane in the feature vector space; this defines a boundary that maximizes the margin between data samples in different classes, therefore leading to good generalization properties. A key factor in the SVM is to use kernels to construct nonlinear decision boundary. In the present work, linear kernels are used. The source code of the SVM is downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm.

2) K-NN Rule: The K-NN rule [6] is used for evaluating the effectiveness of the reduced feature set for classification. It classifies samples based on closest training samples in the feature space. A sample is classified by a majority vote of its K-neighbors, with the sample being assigned to the class most common amongst its K-nearest neighbors. The value of K, chosen for the K-NN rule, is the square root of number of samples in training set.

3) NB Classifier: The NB classifier [6] is one of the oldest classifiers. It is obtained by using the Bayes rule and assuming features or variables are independent of each other given its class. For the jth sample x_j with m gene expression levels $\{\xi_{1j}, \cdots, \xi_{ij}, \cdots, \xi_{mj}\}$ for the m genes, the posterior probability that x_j belongs to class c is

$$p(c|x_j) \propto \prod_{i=1}^{m} p(\xi_{ij}|c) \tag{3}$$

where $p(\xi_{ij}|c)$ are conditional tables or conditional density estimated from training examples.

C. Optimum Value of Threshold

The threshold δ plays an important role to form the gene cluster. It controls the degree of similarity among the genes of a cluster. In effect, it has a direct influence on the performance

TABLE I
COMPARATIVE PERFORMANCE ANALYSIS OF DIFFERENT METHODS USING LOOCV ON SIX CANCER MICROARRAY DATA SETS

Different Data Sets	Different Measures	Algorithms / $d = 1$				Algorithms / $d = 2$				Algorithms / $d = 3$				Full Gene
		MSG	SGCA	ACA	mRMR	MSG	SGCA	ACA	mRMR	MSG	SGCA	ACA	mRMR	
Breast I	SVM	100	100	81.6	85.7	100	100	81.6	81.6	100	100	89.8	93.9	91.8
	K-NN	100	100	81.6	89.8	100	100	83.7	89.8	100	100	83.7	95.9	73.5
	NB	100	100	81.6	89.8	100	100	81.6	95.9	100	100	83.7	100	51.0
Leukemia	SVM	100	93.1	82.4	90.3	100	94.4	88.2	94.4	100	94.4	88.2	94.4	98.6
	K-NN	100	94.4	82.4	93.1	100	94.4	82.4	94.4	100	95.8	91.2	95.8	76.4
	NB	100	94.4	88.2	94.4	100	94.4	88.2	98.6	100	94.4	88.2	100	65.3
Colon	SVM	100	72.6	72.6	83.9	96.8	75.8	72.6	83.9	*	77.4	77.4	75.8	82.3
	K-NN	100	72.6	77.4	83.9	93.6	77.4	83.9	83.9	*	77.4	83.9	83.9	74.2
	NB	100	75.8	64.5	83.9	98.4	77.4	75.8	83.9	*	77.4	64.5	83.9	64.5
Lung	SVM	100	96.7	82.9	97.8	100	96.7	82.9	96.7	100	100	88.4	98.9	98.9
	K-NN	98.9	96.7	82.9	97.8	100	97.2	80.7	96.7	99.5	96.7	85.1	96.7	87.9
	NB	100	80.7	65.2	97.8	100	96.1	80.7	98.9	100	96.1	82.9	98.9	57.1
Prostate	SVM	91.9	82.4	71.3	71.3	91.9	82.4	74.3	82.4	97.8	86.0	74.3	88.2	91.9
	K-NN	94.9	82.4	71.3	80.9	98.5	86.0	80.9	86.0	100	86.0	80.9	86.8	74.2
	NB	97.1	75.7	50.7	84.6	97.8	78.7	55.1	91.2	98.5	78.7	52.9	93.4	56.6
Breast II	SVM	100	80.4	53.6	71.1	100	80.4	53.6	71.1	99.0	89.7	56.7	73.2	68.0
	K-NN	100	80.4	53.6	73.2	100	83.5	55.7	72.2	99.0	89.7	54.6	71.1	63.9
	NB	100	56.7	54.6	75.3	100	60.8	60.8	80.4	100	78.4	60.8	80.4	57.4

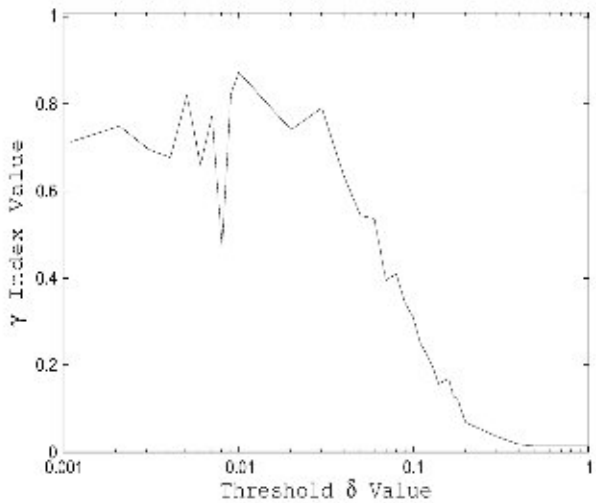

Fig. 2. Variation of γ index for different values of threshold δ for Breast II.

of the proposed supervised gene clustering algorithm. If δ increases, the number of genes in a cluster decreases, but the similarity among them increases. On the other hand, the similarity among the genes of a cluster decreases with the decrease in the value of δ. To find out the optimum value of δ, the γ index is introduced next that is defined as follows:

$$\gamma = \prod_{i=1}^{d} \mathrm{R}_{\mathcal{X}_i}(\mathrm{D}); \qquad 0 \le \gamma \le 1 \tag{4}$$

where $\mathrm{R}_{\mathcal{X}_i}(\mathrm{D})$ represents the relevance value of augmented representative $\mathcal{X}_i$ with respect to the class label D and d is the number of selected representatives.

For six cancer microarray data sets, the value of δ is varied from 0.001 to 1.0 and the corresponding γ index is computed. Fig. 2 represents the variation of the γ index with respect to different values of threshold δ on Breast II data set. From the results reported in Fig. 2, it is seen that as the threshold δ increases, the γ index increases and attains its maximum value at a particular value of δ. After that the γ index decreases with the increase in the value of δ. Hence, the optimum value of δ for each data set is obtained using the following relation:

$$\delta_{\text{optimum}} = \arg\max_{\delta}\{\gamma\}. \tag{5}$$

The optimum values of δ obtained using (5) are 0.15, 0.13, 0.02, 0.17, 0.19, and 0.01 for Breast I, Leukemia, Colon, Prostate, Lung, and Breast II data sets, respectively. On the other hand, the threshold β is used to decide whether a gene of the current cluster would be considered for next cluster generation step or not. From extensive experimentation, the value of β is set to 90% of the maximum similarity of initial cluster representative for each cluster of all the data sets.

D. Comparative Performance Analysis

Table I compares the performance of the proposed MSG algorithm with the best performance of some existing algorithms, namely, ACA [7], SGCA [12], and mRMR [17]. The results are presented based on the classification accuracy of the SVM, K-NN rule, and NB classifier obtained using the LOOCV on six cancer microarray data sets. From the results reported in Table I, it is seen that the proposed algorithm generates a set of clusters having highest classification accuracy of the SVM, K-NN, and NB classifier for all the data sets.

The classification accuracy reported in Table I obtained using the LOOCV is nearly unbiased. However, it is highly variable in the sense that it yields a subset of genes, from the large number of available genes, that has at most only a few genes in common with the subset selected during the original training of the classifier [25]. Hence, the so-called .632+ bootstrap approach [26] is used to reduce the variability of the LOOCV, which is defined as follows:

$$B.632+ = (1 - \omega)E + \omega B \tag{6}$$

where E denotes the proportion of the original training samples misclassified, termed as apparent error rate, and B is the bootstrap error, defined as follows:

$$B = \sum_{j=1}^{n}\sum_{k=1}^{\mathrm{M}} I_{jk}Q_{jk}/n\sum_{k=1}^{\mathrm{M}} I_{jk} \tag{7}$$

where n is the number of original samples and M is the number of bootstrap samples. If the sample x_j is not contained in the kth bootstrap sample, then $I_{jk} = 1$, otherwise 0. Similarly, if x_j is

TABLE II
COMPARATIVE PERFORMANCE ANALYSIS OF DIFFERENT METHODS USING BOOTSTRAP APPROACH ON SIX CANCER MICROARRAY DATA SETS

Different Data Sets	Different Measures	Algorithms / $d = 1$				Algorithms / $d = 2$				Algorithms / $d = 3$				Full Gene
		MSG	SGCA	ACA	mRMR	MSG	SGCA	ACA	mRMR	MSG	SGCA	ACA	mRMR	
Breast I	SVM	97.3	97.3	78.9	82.6	97.3	97.3	81.6	81.6	98.3	98.3	87.1	89.8	87.1
	K-NN	97.3	95.9	78.9	83.7	97.3	95.9	83.7	83.7	98.7	98.5	87.1	89.8	68.3
	NB	98.4	97.9	80.0	81.2	99.1	97.7	82.8	91.2	99.4	98.0	81.6	96.1	52.7
Leukemia	SVM	98.4	90.7	80.5	89.0	98.4	91.1	85.9	92.0	99.7	93.7	85.6	90.1	94.4
	K-NN	99.1	91.4	80.1	90.4	99.7	91.3	80.1	91.7	99.8	94.6	89.8	92.0	65.3
	NB	99.1	90.3	81.5	90.1	99.5	91.8	85.0	93.5	99.5	92.7	86.1	97.8	65.3
Colon	SVM	93.2	71.6	67.2	80.1	94.2	76.3	71.1	80.2	*	76.0	75.1	70.4	83.9
	K-NN	87.6	71.6	69.4	80.1	89.4	72.3	81.5	82.8	*	75.2	81.7	80.6	75.8
	NB	92.1	70.3	63.3	79.6	93.5	75.1	72.7	81.0	*	73.2	61.9	81.6	60.5
Lung	SVM	96.9	93.8	80.1	94.5	98.1	94.1	80.1	92.8	99.0	99.0	83.7	96.0	96.7
	K-NN	97.3	93.8	80.1	94.8	97.4	96.6	81.5	93.2	99.5	95.1	82.8	93.2	85.1
	NB	96.8	82.7	64.9	94.5	97.0	96.3	80.1	95.7	97.0	97.2	80.3	96.9	55.0
Prostate	SVM	90.1	82.5	67.9	69.7	88.1	80.3	70.9	81.0	93.4	82.7	71.6	86.1	91.9
	K-NN	93.8	82.5	67.1	81.1	94.8	83.7	80.2	85.1	99.0	82.8	81.0	83.7	71.3
	NB	94.0	73.1	52.8	80.2	93.3	75.5	56.8	89.0	97.3	76.1	55.7	92.1	52.9
Breast II	SVM	96.4	80.9	54.8	70.9	96.8	77.1	50.2	72.9	97.5	84.2	59.4	71.0	53.6
	K-NN	97.7	80.9	54.8	72.1	98.9	80.6	53.2	70.7	99.1	85.2	55.0	70.9	54.6
	NB	96.4	57.8	53.1	70.6	99.1	61.4	58.4	81.2	97.2	74.9	57.3	78.1	57.4

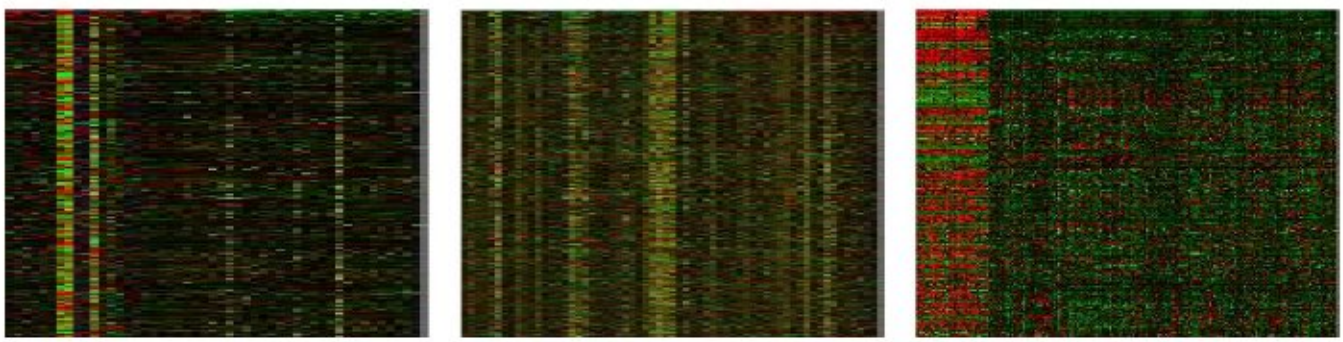

Fig. 3. Eisen plots of the best clusters obtained using proposed algorithm: (a) Breast I: $\delta = 0.15$; (b) Colon: $\delta = 0.02$; (c) Lung: $\delta = 0.19$.

misclassified, $Q_{jk} = 1$, otherwise 0. The weight parameter ω is given by

$$\omega = \frac{0.632}{1 - 0.368r}; \quad r = \frac{B - E}{\gamma - E}; \quad \text{and} \quad \gamma = \sum_{i=1}^{K} p_i(1 - q_i) \tag{8}$$

where p_i is the proportion of the samples from the ith class, q_i is the proportion of them assigned to the ith class, and γ is the no-information error rate. Table II reports the classification accuracy of different algorithms using the relation $(1 - B.632+) \times 100$ based on the .632+ bootstrap approach considering $M = 50$. From the results reported in Table II, it is also seen that the proposed algorithm generates a set of clusters having lowest bootstrap error of the SVM, K-NN rule, and NB classifier for all the data sets.

The better performance of the proposed algorithm is achieved due to the fact that it uses mutual information for computing both gene-class relevance and gene-gene similarity to generate co-regulated gene clusters with strong association to the class labels. As mutual information depends only on the probability distribution of a gene rather than on its actual values, it is insensitive to noise or outlier of the data set. Moreover, the cluster representatives of the proposed algorithm are modified based on the information of class labels. In effect, it can identify functional groups of genes present in microarray data more accurately than existing algorithms. The coherent average expression levels of these functionally similar gene clusters allow perfect discrimination of class labels. Furthermore, Fig. 3 presents the Eisen plots [27] of best clusters for three cancer data sets generated by the proposed algorithm. All the results reported here establish the fact that the proposed algorithm can efficiently identify groups of co-regulated genes with strong association to the sample categories.

The classification accuracy of the SVM, K-NN, and NB for full gene set is also reported in Tables I and II. The results reported in these tables indicate that if all genes are considered for sample classification, the samples from different classes may not be well separated with respect to these three classifiers. However, when a gene selection or clustering algorithm selects a set of genes from full gene set considering the relevance or redundancy criteria, the genes those have high relevance with respect to the class labels are only selected. In effect, the samples from different classes with reduced gene set become well separated, which leads to higher classification accuracy. On the other hand, the presence of irrelevant and redundant genes in reduced gene set may degrade the quality of solution. From the results reported in Table I, it is seen that the classification accuracy of three classifiers obtained using the proposed algorithm and that of the K-NN rule and NB classifier obtained using the mRMR method, ACA, and SGCA is always higher than that achieved by the full gene set for all microarray data sets. On the other hand, out of 6 data sets, the mRMR method and SGCA

TABLE III
SIGNIFICANT SHARED GO TERMS FOR GENES IN BEST CLUSTERS

Ontology	Algorithm	Information	Breast Cancer I	Leukemia	Colon Cancer
Biological Processes	Proposed	Term	Cellular process	Multicellular organismal process	Cellular process
		P-Value	6.57E-078	1.06E-07	1.19E-80
		FDR Rate	0.00%	0.00%	0.00%
		False Positives	0.00	0.00	0.00
	SGCA	Term	*	*	Blood vessel morphogenesis
		P-Value	*	*	2.22E-02
		FDR Rate	*	*	30.00%
		False Positives	*	*	0.30
Molecular Functions	Proposed	Term	Binding	Binding	Protein binding
		P-Value	1.70E-069	1.08E-06	2.42E-78
		FDR Rate	0.00%	0.00%	0.00%
		False Positives	0.00	0.00	0.00
	SGCA	Term	*	Structural constituent of muscle	*
		P-Value	*	1.83E-03	*
		FDR Rate	*	8.00%	*
		False Positives	*	0.08	*
Cellular Components	Proposed	Term	Intracellular part	Cytoplasmic part	Intracellular part
		P-Value	5.81E-053	2.46E-07	2.32E-77
		FDR Rate	0.00%	0.00%	0.00%
		False Positives	0.00	0.00	0.00
	SGCA	Term	*	*	Cytoplasm
		P-Value	*	*	4.66E-02
		FDR Rate	*	*	6.00%
		False Positives	*	*	0.06

perform better than full gene set in 3 cases with respect to the SVM. Similar results can also be found in case of Table II.

E. Biological Significance Analysis

To interpret the biological significance of the generated clusters, the Gene Ontology (GO) Term Finder is used [28]. It finds the most significantly enriched GO terms associated with the genes belonging to a cluster. The GO project aims to build tree structures, controlled vocabularies, also called ontologies, that describe gene products in terms of their associated biological processes (BP), molecular functions (MF) or cellular components (CC). The GO Term Finder determines whether any GO term annotates a specified list of genes at a frequency greater than that would be expected by chance, calculating the associated p-value by using the hypergeometric distribution and the Bonferroni multiple-hypothesis correction [28]. The closer the p-value is to zero, the more significant the particular GO term associated with the group of genes is, that is, the less likely the observed annotation of the particular GO term to a group of genes occurs by chance. On the other hand, the false discovery rate (FDR) is a multiple-hypothesis testing error measure indicating the expected proportion of false positives among the set of significant results.

Hence, the GO Term Finder is used to determine the statistical significance of the association of a particular GO term with the genes of best cluster produced by the proposed algorithm. The GO Term Finder is used to compute the p-value, FDR (%), and false positives for all the GO terms from the BP, MF, and CC ontology and the most significant term, that is, the one with the lowest p-value, is chosen to represent the set of genes of best cluster. Table III presents the significant shared GO terms, along with the p-values, FDR(%), and false positives for the BP, MF, and CC on different data sets. The results corresponding to the best clusters of the existing SGCA [12] are also provided on same data sets for the sake of comparison. The "*" in Table III represents that no significant shared term is found considering p-value cutoff as 0.05. From the results reported in [15] and Table III, it is seen that the best cluster generated by the proposed algorithm can be assigned to the GO BPs, MFs, and CCs with high reliability in terms of p-value, FDR, and false positives. That is, the proposed algorithm describes accurately the known classification, the one given by the GO, and thus reliable for extracting new biological insights. The annotated genes of best cluster produced by the proposed algorithm for three ontologies, determined by the Go Term Finder, are reported at www.isical.ac.in/~pmaji/results/tnb.html.

IV. CONCLUSION

This paper presents a supervised gene clustering algorithm for cancer classification using microarray experiments. The proposed algorithm is potentially useful in the context of medical diagnostics as it identifies groups of interacting genes that have high explanatory power for given tissue types, and which in turn can accurately predict the class labels of new samples. Moreover, the generated clusters reveal insights into biological processes that may be valuable for functional genomics. In brief, the proposed algorithm tries to cluster genes in such a way that the discrimination of different tissue types becomes as simple as possible. Comparing to existing supervised gene clustering approaches, only the proposed method generates overlapping gene clusters and is also applicable to multiclass classification problem.

The performance of the proposed method is evaluated by the predictive accuracy of the NB classifier, K-NN rule, and SVM. For all data sets, significantly better results are found by the proposed method compared to other methods. The results obtained on real data sets demonstrate that the proposed method can bring a remarkable improvement on gene clustering problem. All the results reported in this paper demonstrate the feasibility and effectiveness of the proposed method. The proposed method is capable of identifying discriminative genes that may contribute to revealing underlying class structures, providing a useful tool for the exploratory analysis of biological data.

ACKNOWLEDGMENT

The authors would like to thank anonymous referees for providing helpful comments and valuable criticisms on the original version of the manuscript.

REFERENCES

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, 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, no. 5439, pp. 531–537, 1999.

[2] E. Domany, "Cluster analysis of gene expression data," *J. Stat. Phys.*, vol. 110, no. 3–6, pp. 1117–1139, 2003.

[3] J. G. Liao and K.-V. Chin, "Logistic regression for disease classification using microarray data: Model selection in a large p and small n case," *Bioinformatics*, vol. 23, no. 15, pp. 1945–1951, 2007.

[4] P. Maji, "f-information measures for efficient selection of discriminative genes from microarray data," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1063–1069, 2009.

[5] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification and Scene Analysis*. New York: Wiley, 1999.

[7] W.-H. Au, K. C. C. Chan, A. K. C. Wong, and Y. Wang, "Attribute clustering for grouping, selection, classification of gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 2, no. 2, pp. 83–101, 2005.

[8] J. Herrero, A. Valencia, and J. Dopazo, "A hierarchical unsupervised growing neural network for clustering gene expression patterns," *Bioinformatics*, vol. 17, pp. 126–136, 2001.

[9] L. J. Heyer, S. Kruglyak, and S. Yooseph, "Exploring expression data: Identification and analysis of coexpressed genes," *Genome Res.*, vol. 9, no. 11, pp. 1106–1115, 1999.

[10] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proc. Natl. Acad. Sci. USA*, vol. 96, no. 6, pp. 2907–2912, 1999.

[11] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1370–1386, 2004.

[12] M. Dettling and P. Buhlmann, "Supervised clustering of genes," *Genome Biology*, vol. 3, no. 12, pp. 1–15, 2002.

[13] T. Hastie, R. Tibshirani, D. Botstein, and P. Brown, "Supervised harvesting of expression trees," *Genome Biol.*, vol. 1, pp. 1–12, 2001.

[14] D. Nguyen and D. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, pp. 39–50, 2002.

[15] P. Maji, "Mutual information based supervised attribute clustering for microarray sample classification," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 1, pp. 127–140, 2012.

[16] C. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Champaign, IL: Univ. Illinois Press, 1964.

[17] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinf. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, 2005.

[18] X. Liu, A. Krishnan, and A. Mondry, "An entropy based gene selection method for cancer classification using microarray data," *BMC Bioinf.*, vol. 6, no. 76, pp. 1–14, 2005.

[19] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[20] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles," *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 20, pp. 11 462–11 467, 2001.

[21] 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 tissues probed by Oligonucleotide arrays," *Proc. Natl. Acad. Sci. USA*, vol. 96, no. 12, pp. 6745–6750, 1999.

[22] G. J. Gordon, R. V. Jensen, 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 mesothelioma," *Cancer Res.*, vol. 62, pp. 4963–4967, 2002.

[23] L. J. v. Veer, H. Dai, M. J. v. D. Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. v. d. Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, 2002.

[24] 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 correlates of clinical prostate cancer behavior," *Cancer Res.*, vol. 1, pp. 203–209, 2002.

[25] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 10, pp. 6562–6566, 2002.

[26] B. Efron and R. Tibshirani, "Improvements on cross-validation: The .632 + bootstrap method," *J. Amer. Stat. Assoc.*, vol. 92, no. 438, pp. 548–560, 1997.

[27] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA*, vol. 95, no. 25, pp. 14 863–14 868, 1998.

[28] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock, "GO::Term finder open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, pp. 3710–3715, 2004.

Pradipta Maji received the B.Sc. degree in physics, the M.Sc. degree in electronics science, and the Ph.D. degree in the area of computer science from Jadavpur University, India, in 1998, 2000, and 2005, respectively.

Currently, he is an Assistant Professor in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata. He has published more than 70 papers in international journals and conferences. He is an author of a book published by Wiley-IEEE Computer Society Press, and also a reviewer of many international journals. His research interests include pattern recognition, machine learning, computational biology and bioinformatics, medical image processing, and so forth.

Dr. Maji has received the 2006 Best Paper Award of the International Conference on Visual Information Engineering from The Institution of Engineering and Technology, U.K., the 2008 Microsoft Young Faculty Award from Microsoft Research Laboratory India Pvt., the 2009 Young Scientist Award from the National Academy of Sciences, India, and the 2011 Young Scientist Award from the Indian National Science Academy, India, and has been selected as the 2009 Young Associate of the Indian Academy of Sciences, India.

Chandra Das received the B.Sc. degree in computer science, the M.Sc. degree in computer and information science, and the M.Tech. degree in computer science and engineering from the University of Calcutta, India, in 1999, 2001, and 2003, respectively, and the Ph.D. degree in computer science from Jadavpur University, India, in 2011.

Currently, she is a Senior Lecturer in the Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India. She has a number of publications in international journals and conferences. Her research interests include pattern recognition, machine learning, computational biology, and bioinformatics.