Mutual Information-Based Supervised Attribute Clustering for Microarray Sample Classification

Pradipta Maji

Abstract—Microarray technology is one of the important biotechnological means that allows to record the expression levels of thousands of genes simultaneously within a number of different samples. An important application of microarray gene expression data in functional genomics is to classify samples according to their gene expression profiles. Among the large amount of genes presented in gene expression data, only a small fraction of them is effective for performing a certain diagnostic test. Hence, one of the major tasks with the gene expression data is to find groups of coregulated genes whose collective expression is strongly associated with the sample categories or response variables. In this regard, a new supervised attribute clustering algorithm is proposed to find such groups of genes. It directly incorporates the information of sample categories into the attribute clustering process. A new quantitative measure, based on mutual information, is introduced that incorporates the information of sample categories to measure the similarity between attributes. The proposed supervised attribute clustering algorithm is based on measuring the similarity between attributes using the new quantitative measure, whereby redundancy among the attributes is removed. The clusters are then refined incrementally based on sample categories. The performance of the proposed algorithm is compared with that of existing supervised and unsupervised gene clustering and gene selection algorithms based on the class separability index and the predictive accuracy of naive bayes classifier, K-nearest neighbor rule, and support vector machine on three cancer and two arthritis microarray data sets. The biological significance of the generated clusters is interpreted using the gene ontology. An important finding is that the proposed supervised attribute clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability.

Index Terms—Microarray analysis, attribute clustering, gene selection, mutual information, classification.

1 Introduction

Recent advancement and wide use of high-throughput technology are producing an explosion in using gene expression phenotype for identification and classification in a variety of diagnostic areas. An important application of gene expression data in functional genomics is to classify samples according to their gene expression profiles [1], [2].

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, for most gene expression data, the number of training samples is still very small compared to the large number of genes involved in the experiments. When the number of genes is significantly greater than the number of samples, it is possible to find biologically relevant correlations of gene behavior with the sample categories or response variables [3].

However, among the large amount of genes, only a small fraction is effective for performing a certain task. Also, a small subset of genes is desirable in developing gene expression-based diagnostic tools for delivering precise, reliable, and interpretable results [4]. With the gene selection results, the cost of biological experiment and decision can be greatly reduced by analyzing only the marker genes. Hence, identifying a reduced set of most relevant genes is the goal of gene selection. The small number of training samples and a large number of genes make gene selection a more relevant and challenging problem in gene expression-based classification. As this is a feature selection problem [5], [6], [7], the clustering method can be used, which partitions the given gene set into subgroups, each of which should be as homogeneous as possible [8], [9], [10], [11], [12].

When applied to gene expression data analysis, the conventional clustering methods such as bayesian clustering [13], [14], hierarchical clustering [15], [16], k-means algorithm [17], self-organizing map [16], [18], and principal component analysis [19], [20] group a subset of genes that are interdependent or correlated with each other. In other words, genes or attributes in a cluster are more correlated with each other, whereas genes in different clusters are less correlated [10], [11], [12]. The attribute clustering is able to reduce the 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 [10], [11], [12].

However, all these algorithms group genes according to unsupervised similarity measures computed from the gene expressions, without using any information about the sample categories or response variables. The information of response variables should be incorporated in attribute clustering to find groups of coregulated genes with strong association to the sample categories [21]. In this background, some supervised attribute clustering algorithms

- *The author is with the Machine Intelligence Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700 108, West Bengal, India. E-mail: pmaji@isical.ac.in.*

such as supervised gene clustering [21], gene shaving [22], tree harvesting [23], and partial least square procedure [24] have been proposed to reveal groups of coregulated genes with strong association to the sample categories. The supervised attribute clustering is defined as the grouping of genes or attributes, controlled by the information of sample categories or response variables.

In general, the quality of generated clusters is always relative to a certain criterion. Different criteria may lead to different clustering results. However, every criterion tries to measure the similarity among the subset of genes presented in a cluster. While tree harvesting [23] uses an unsupervised similarity measure to group a set of coregulated genes, other supervised algorithms such as supervised gene clustering [21], gene shaving [22], and partial least square procedure [24] do not use any similarity measure to cluster genes; rather use different predictive scores such as Wilcoxon test [21] and Cox model score test [22] to measure gene-class relevance. Moreover, all these measures depend on the actual values of the training data. Hence, they may be sensitive to noise or outlier of the data set [10], [17], [25], [26]. On the other hand, as mutual information [25], [26], [27], [28] depends only on the probability distribution of a random variable, it has been widely used for computing both gene-class relevance and gene-gene redundancy or similarity [11], [25], [26], [27], [28], [29], [30].

In this regard, a new supervised attribute clustering algorithm is proposed to find coregulated clusters of genes whose collective expression is strongly associated with the sample categories or class labels. A new quantitative measure, based on mutual information, is introduced to compute the similarity between attributes. The proposed measure incorporates the information of sample categories while measuring the similarity between attributes. In effect, it helps to identify functional groups of genes that are of special interest in sample classification. The proposed supervised attribute clustering method 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 coregulated 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. Finally, the modified representative of each cluster is selected to constitute the resulting reduced feature set. In effect, the proposed supervised attribute clustering algorithm yields biologically significant gene clusters, whose coherent average expression levels allow perfect discrimination of sample categories. Also, the proposed algorithm avoids the noise sensitivity problem of existing supervised gene clustering algorithms. The performance of the proposed algorithm, along with a comparison with existing algorithms is studied both qualitatively and quantitatively on three cancer and two arthritis data sets using the class separability index and the predictive accuracy of naive bayes (NB) classifier, K-nearest neighbor rule (K-NN), and support vector machine.

The structure of the rest of this paper is as follows: Section 2 briefly introduces existing supervised and unsupervised gene clustering algorithms, along with different existing criteria used for computing the relevance and redundancy. The proposed supervised attribute clustering algorithm is presented in Section 3. A few case studies and a comparison with existing algorithms are presented in Section 4. Concluding remarks are given in Section 5.

2 Clustering of Gene Expression Data

In this section, some existing supervised and unsupervised gene clustering algorithms are reported, along with different widely used criteria for computing gene-class relevance and gene-gene redundancy.

2.1 Gene Clustering

Clustering is one of the major tasks in gene expression data analysis. To find groups of coregulated genes from microarray data, different unsupervised clustering techniques such as hierarchical clustering [15], [16], k-means algorithm [17], self-organizing map [16], [18], and principal component analysis [19], [20] have been widely used. The hierarchical clustering identifies sets of correlated genes with similar behavior across the samples, but yields thousands of clusters in a tree-like structure, which makes the identification of functional groups very difficult [15], [16]. In contrast, self-organizing map [16], [18] and k-means algorithm [17] require a prespecified number and an initial spatial structure of clusters, but this may be hard to come up with in real problems. 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 or class labels [21].

To reveal groups of coregulated genes with strong association to the sample categories, different supervised attribute clustering algorithms have been proposed recently [21], [22], [23], [24]. One notable work in this field encompasses tree harvesting [23], 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 the tissue types. Another supervised clustering method, called gene shaving, identifies subsets of genes with coherent expression patterns and large variation across the conditions [22]. The technique can be unsupervised, where the genes and samples are treated as unlabeled, or partially or fully supervised by using known properties of the genes or samples to assist in finding meaningful groupings.

An interesting supervised clustering approach that directly incorporates the response variables in the grouping process is the partial least squares procedure [24], which in a supervised 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 [21]. A similar single step approach is also pursued by Jornsten and Yu [31]. The supervised attribute clustering algorithm proposed in [21] 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 [21], 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.

2.2 Criteria for Gene Selection and Clustering

The t-test, F-test [25], [32], information gain, mutual information [25], [26], normalized mutual information [29], and f-information [33] are typically used to measure the relevance of a gene with respect to the class labels or sample categories and the same or a different metric such as mutual information, the L_1 distance, euclidean distance, and Pearson's correlation coefficient [10], [25], [26] is employed to calculate the similarity or redundancy between genes.

To measure the relevance of a gene, the t-test is widely used, assuming that there are two classes of samples in a gene expression data set. When there are multiple classes of samples, the t-test is typically computed for one class versus all the other classes. For multiple classes of samples, an F-test between a gene and the class label can be used to calculate the relevance score of that gene. The F-test reduces to the t-test for two class problem with the relation $F = t^2$. In [21], the Wilcoxon's test statistic is used to compute the relevance of a gene assuming two classes of samples in microarray data set.

On the other hand, the euclidean distance measures the difference in the individual magnitudes of each gene. However, the genes regarded as similar by the euclidean distance may be very dissimilar in terms of their shapes. Similarly, the euclidean distance between two genes having an identical shape may be large if they differ from each other by a large scaling factor. But, the overall shapes of genes are of the primary interest for gene expression data [10]. Hence, the euclidean distance may not be able to yield a good proximity measurement of genes [10]. The Pearson's correlation coefficient considers each gene as a random variable and measures the similarity between two genes by calculating the linear relationship between distributions of two corresponding random variables. An empirical study has shown that Pearson's correlation coefficient is not robust to outliers and it may assign high similarity score to a pair of dissimilar genes [17].

However, as the t-test, F-test, Wilcoxon's test, euclidean distance, and Pearson's correlation depend on the actual gene expression values of microarray data, they are very much sensitive to noise or outlier of the data set. On the other hand, as the information theoretic measure such as entropy, mutual information, and f-information depends only on the probability distribution of a random variable rather than on its actual values, it is more effective to evaluate the gene-class relevance as well as gene-gene redundancy [25], [26], [33].

In principle, the mutual information is used to quantify the information shared by two objects. If two independent objects do not share much information, the mutual information value between them is small. While two highly correlated objects will demonstrate a high mutual information value [34]. The objects can be the class label and the genes. The necessity for a gene to be an independent and informative can, therefore, be determined by the shared information between the gene and the rest as well as the shared information between the gene and class label [25], [26]. If a gene has expression values randomly or uniformly 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. Thus, the mutual information can be used as a measure of relevance of genes. Similarly, the mutual information may be used to measure the level of similarity or redundancy between two genes.

3 Proposed Clustering Algorithm

In this section, a new supervised attribute clustering algorithm is presented for grouping coregulated genes with strong association to the class labels. It is based on a supervised similarity measure that follows next.

3.1 Supervised Similarity Measure

In real data analysis, one of the important issues is computing both relevance and redundancy of attributes by discovering dependencies among them. Intuitively, a set of attributes $\mathbb{Q}$ depends totally on a set of attributes $\mathbb{P}$, if all attribute values from $\mathbb{Q}$ are uniquely determined by values of attributes from $\mathbb{P}$. If there exists a functional dependency between values of $\mathbb{Q}$ and $\mathbb{P}$, then $\mathbb{Q}$ depends totally on $\mathbb{P}$.

Let $\mathbb{U} = \{x_1, \ldots, x_i, \ldots, x_n\}$ is the set of n samples and $\mathbb{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots, \mathcal{A}_j, \ldots, \mathcal{A}_m\}$ denotes the set of m attributes of a given data set $\mathcal{T} = \{w_{ij} | i = 1, \ldots, m, j = 1, \ldots, n\}$, where $w_{ij} \in \Re$ is the measured value of the attribute $\mathcal{A}_i$ in the sample x_j. Let $\mathbb{D} = \{\mathrm{D}_1, \ldots, \mathrm{D}_i, \ldots, \mathrm{D}_n\}$ represents the set of class labels or sample categories of n samples. Define $\mathrm{R}_{\mathcal{A}_i}(\mathbb{D})$ as the relevance of the attribute $\mathcal{A}_i$ with respect to the class label $\mathbb{D}$ while $\mathrm{S}(\mathcal{A}_i, \mathcal{A}_j)$ as the redundancy or similarity between two attributes $\mathcal{A}_i$ and $\mathcal{A}_j$. The mutual information can be used to calculate both relevance and redundancy among attributes.

The relevance $\mathrm{R}_{\mathcal{A}_i}(\mathbb{D})$ of the attribute $\mathcal{A}_i$ with respect to the class label $\mathbb{D}$ using mutual information can be calculated as follows:

$$\mathrm{R}_{\mathcal{A}_i}(\mathbb{D}) = I(\mathcal{A}_i, \mathbb{D}), \tag{1}$$

where $I(\mathcal{A}_i, \mathbb{D})$ represents the mutual information between attribute $\mathcal{A}_i$ and class label $\mathbb{D}$ that is given by

$$I(\mathcal{A}_i, \mathbb{D}) = H(\mathcal{A}_i) - H(\mathcal{A}_i|\mathbb{D}). \tag{2}$$

Here, $H(\mathcal{A}_i)$ and $H(\mathcal{A}_i|\mathbb{D})$ represent the entropy of attribute $\mathcal{A}_i$ and the conditional entropy of $\mathcal{A}_i$ given class label $\mathbb{D}$, respectively. The entropy $H(\mathcal{A}_i)$ is known to be a measure of the amount of uncertainty about the attribute $\mathcal{A}_i$ while $H(\mathcal{A}_i|\mathbb{D})$ is the amount of uncertainty left in $\mathcal{A}_i$ when knowing $\mathbb{D}$. Hence, the quantity $I(\mathcal{A}_i, \mathbb{D})$ is the reduction in the uncertainty of the attribute $\mathcal{A}_i$ by the knowledge of class label $\mathbb{D}$. In other words, it represents the amount of information that the class label $\mathbb{D}$ contains about the attribute $\mathcal{A}_i$.

Definition 1. *For continuous random variables such as gene expression values, the entropy, conditional entropy, and mutual information can be defined as follows:*

$$H(\mathcal{Y}) = -\int p(y) \log p(y) dy, \tag{3}$$

$$H(\mathcal{Y}|\mathcal{Z}) = -\int p(y, z) \log p(y|z) dy dz, \tag{4}$$

$$I(\mathcal{Y}, \mathcal{Z}) = \iint p(y, z) \log \frac{p(y, z)}{p(y)p(z)} dy dz, \tag{5}$$

where $p(y)$ is the true probability density function of the attribute or variable $\mathcal{Y}$, while $p(y|z)$ and $p(y, z)$ represent the conditional probability density function of $\mathcal{Y}$ given the variable $\mathcal{Z}$ and the joint probability density function of $\mathcal{Y}$ and $\mathcal{Z}$, respectively. Usually, the Gaussian function is used to approximate the true density function [35].

The redundancy or similarly between two attributes $\mathcal{A}_i$ and $\mathcal{A}_j$, in terms of mutual information, can also be calculated as follows:

$$\mathrm{S}(\mathcal{A}_i, \mathcal{A}_j) = I(\mathcal{A}_i, \mathcal{A}_j). \tag{6}$$

However, the term $\mathrm{S}(\mathcal{A}_i, \mathcal{A}_j)$ does not incorporate the information of sample categories or class labels $\mathbb{D}$ while measuring the similarity and it is considered as unsupervised similarity measure. Hence, a new quantitative measure, called supervised similarity measure is defined here based on mutual information for measuring the similarity between two random variables. It incorporates the information of sample categories or class labels while measuring the similarity between attributes.

Definition 2. *The significance of an attribute $\mathcal{A}_j$ with respect to another attribute $\mathcal{A}_i$ can be defined as follows:*

$$\sigma_{\mathcal{A}_i}(\mathcal{A}_j) = \mathrm{R}_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}) - \mathrm{R}_{\mathcal{A}_i}(\mathbb{D}). \tag{7}$$

That is, the significance of an attribute $\mathcal{A}_j$ is the change in dependency when the attribute $\mathcal{A}_j$ is removed from the set $\{\mathcal{A}_i, \mathcal{A}_j\}$. The higher the change in dependency, the more significant the attribute $\mathcal{A}_j$ is. If the significance is 0, then the attribute $\mathcal{A}_j$ is dispensable.

Based on the concept of significance of an attribute, the supervised similarity measure between two attributes is defined next.

Definition 3. *The supervised similarity between two attributes $\mathcal{A}_i$ and $\mathcal{A}_j$ is defined as follows:*

$$\Psi(\mathcal{A}_i, \mathcal{A}_j) = \frac{1}{1 + \kappa^2}, \tag{8}$$

$$\text{where } \kappa = \left\{ \frac{\sigma_{\mathcal{A}_i}(\mathcal{A}_j) + \sigma_{\mathcal{A}_j}(\mathcal{A}_i)}{2} \right\}, \tag{9}$$

$$\text{that is, } \kappa = \mathrm{R}_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}) - \left\{ \frac{\mathrm{R}_{\mathcal{A}_i}(\mathbb{D}) + \mathrm{R}_{\mathcal{A}_j}(\mathbb{D})}{2} \right\}. \tag{10}$$

Hence, the supervised similarity measure $\Psi(\mathcal{A}_i, \mathcal{A}_j)$ directly takes into account the information of sample categories or class labels $\mathbb{D}$ while computing the similarity between two attributes $\mathcal{A}_i$ and $\mathcal{A}_j$. If attributes $\mathcal{A}_i$ and $\mathcal{A}_j$ are completely correlated with respect to class labels $\mathbb{D}$, then $\kappa = 0$ and so $\Psi(\mathcal{A}_i, \mathcal{A}_j)$ is 1. If $\mathcal{A}_i$ and $\mathcal{A}_j$ are totally uncorrelated, $\Psi(\mathcal{A}_i, \mathcal{A}_j) \rightarrow 0$. Hence, $\Psi(\mathcal{A}_i, \mathcal{A}_j)$ can be used as a measure of supervised similarity between two attributes $\mathcal{A}_i$ and $\mathcal{A}_j$. The following properties can be stated about the measure:

1. $0 < \Psi(\mathcal{A}_i, \mathcal{A}_j) \leq 1$.
2. $\Psi(\mathcal{A}_i, \mathcal{A}_j) = 1$ if and only if $\mathcal{A}_i$ and $\mathcal{A}_j$ are completely correlated.
3. $\Psi(\mathcal{A}_i, \mathcal{A}_j) \rightarrow 0$ if and only if $\mathcal{A}_i$ and $\mathcal{A}_j$ are totally uncorrelated.
4. $\Psi(\mathcal{A}_i, \mathcal{A}_j) = \Psi(\mathcal{A}_j, \mathcal{A}_i)$ (symmetric).

The supervised similarity between two attributes $\mathcal{A}_i$ and $\mathcal{A}_j$, in terms of entropy is given by

$$\Psi(\mathcal{A}_i, \mathcal{A}_j) = \left[1 + \left[H(\mathcal{A}_i\mathcal{A}_j|\mathbb{D}) - \frac{1}{2}\{ H(\mathcal{A}_i|\mathcal{A}_j) + H(\mathcal{A}_j|\mathcal{A}_i) + H(\mathcal{A}_i|\mathbb{D}) + H(\mathcal{A}_j|\mathbb{D}) \} \right]^2 \right]^{-1}. \tag{11}$$

Combining (6) and (11), the term $\Psi(\mathcal{A}_i, \mathcal{A}_j)$ can be expressed as follows:

$$\Psi(\mathcal{A}_i, \mathcal{A}_j) = \left[1 + \left[\mathrm{S}(\mathcal{A}_i, \mathcal{A}_j) + H(\mathcal{A}_i\mathcal{A}_j|\mathbb{D}) - \frac{1}{2}\{ H(\mathcal{A}_i) + H(\mathcal{A}_j) + H(\mathcal{A}_i|\mathbb{D}) + H(\mathcal{A}_j|\mathbb{D}) \} \right]^2 \right]^{-1}. \tag{12}$$

Hence, the supervised similarity measure $\Psi(\mathcal{A}_i, \mathcal{A}_j)$ not only considers the information of sample categories or class labels $\mathbb{D}$, it also takes into account the unsupervised similarity between two attributes $\mathrm{S}(\mathcal{A}_i, \mathcal{A}_j)$.

3.2 Supervised Attribute Clustering Algorithm

The proposed supervised attribute clustering algorithm relies on mainly two factors, namely, determining the relevance of each attribute and growing the cluster around each relevant attribute incrementally by adding one attribute after the other. One of the important property of the proposed clustering approach is that the cluster is augmented by the attributes those satisfy following two conditions:

1. Suit best into the current cluster in terms of a supervised similarity measure defined above.
2. Improve the differential expression of the current cluster most, according to the relevance of the cluster representative or prototype.

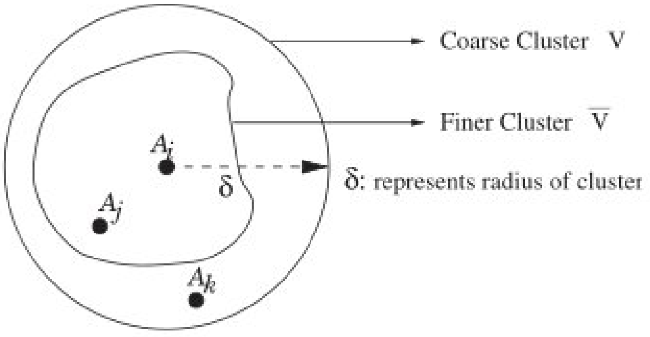

Coarse Cluster $\mathbb{V}$
Finer Cluster $\bar{\mathbb{V}}$
$\mathcal{A}_i$
δ
δ: represents radius of cluster
$\mathcal{A}_j$
$\mathcal{A}_k$

Fig. 1. Representation of a supervised attribute cluster.

The growth of a cluster is repeated until the cluster stabilizes, and then the proposed clustering algorithm starts to generate a new cluster.

Let $\mathrm{R}_{\mathcal{A}_i}(\mathbb{D})$ represents the relevance of attribute $\mathcal{A}_i \in \mathbb{A}$ with respect to 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 attribute $\mathcal{A}_i$ that has the highest relevance value with respect to class labels. An initial cluster $\mathbb{V}_i$ is formed by selecting the set of attributes $\{\mathcal{A}_j\}$ from the whole set $\mathbb{A}$ considering the attribute $\mathcal{A}_i$ as the representative of cluster $\mathbb{V}_i$, where

$$\mathbb{V}_i = \{\mathcal{A}_j | \Psi(\mathcal{A}_i, \mathcal{A}_j) \geq \delta; \mathcal{A}_j \neq \mathcal{A}_i \in \mathbb{A}\}. \tag{13}$$

Hence, the cluster $\mathbb{V}_i$ represents the set of attributes of $\mathbb{A}$ those have the supervised similarity values with the attribute $\mathcal{A}_i$ greater than a predefined threshold value δ. The cluster $\mathbb{V}_i$ is the coarse cluster corresponding to the attribute $\mathcal{A}_i$, while the threshold δ is termed as the radius of cluster $\mathbb{V}_i$ (Fig. 1).

After forming the initial coarse cluster $\mathbb{V}_i$, the cluster representative is refined incrementally. By searching among the attributes of cluster $\mathbb{V}_i$, the current cluster representative is merged and averaged with one single attribute such that the augmented cluster representative $\bar{\mathcal{A}}_i$ increases the relevance value. The merging process is repeated until the relevance value can no longer be improved. Instead of averaging all attributes of $\mathbb{V}_i$, the augmented attribute $\bar{\mathcal{A}}_i$ is computed by considering a subset of attributes $\bar{\mathbb{V}}_i \subset \mathbb{V}_i$ those increase the relevance value of cluster representative $\bar{\mathcal{A}}_i$. The set of attributes $\bar{\mathbb{V}}_i$ represents the finer cluster of the attribute $\mathcal{A}_i$ (Fig. 1). While the generation of coarse cluster reduces the redundancy among attributes of the set $\mathbb{A}$, that of finer cluster increases the relevance with respect to class labels. After generating the augmented cluster representative $\bar{\mathcal{A}}_i$ from the finer cluster $\bar{\mathbb{V}}_i$, the process is repeated to find more clusters and augmented cluster representatives by discarding the set of attributes $\mathbb{V}_i$ from the whole set $\mathbb{A}$.

To compute the set $\mathbb{V}_i$ corresponding to the attribute $\mathcal{A}_i$, one may consider the conventional unsupervised similarity measure $\mathrm{S}(\mathcal{A}_i, \mathcal{A}_j)$ as defined in (6). However, as it does not take into account the information of sample categories or class labels, the attributes are clustered by similarity only, without using any information about the sample categories. In effect, it fails to reveal functional groups of attributes that are of special interest in classification. On the other hand, as the supervised similarity measure $\Psi(\mathcal{A}_i, \mathcal{A}_j)$ defined in (8) incorporates the class information directly while computing the similarity between two attributes $\mathcal{A}_i$ and $\mathcal{A}_j$, it can identify functional groups present in the attribute set.

The main steps of the proposed supervised attribute clustering algorithm are reported next.

- Let $\mathbb{C}$ represents the set of attributes of the original data set, while $\mathbb{S}$ and $\bar{\mathbb{S}}$ are the set of actual and augmented attributes, respectively, selected by the proposed attribute clustering algorithm.
- Let $\mathbb{V}_i$ is the coarse cluster associated with the attribute $\mathcal{A}_i$ and $\bar{\mathbb{V}}_i$, the finer cluster of $\mathcal{A}_i$ (Fig. 1), represents the set of attributes of $\mathbb{V}_i$ those are merged and averaged with the attribute $\mathcal{A}_i$ to generate the augmented cluster representative $\bar{\mathcal{A}}_i$.

1. Initialize $\mathbb{C} \leftarrow \mathbb{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots, \mathcal{A}_j, \ldots, \mathcal{A}_m\}$, $\mathbb{S} \leftarrow \emptyset$, and $\bar{\mathbb{S}} \leftarrow \emptyset$.
2. Calculate the relevance value $\mathrm{R}_{\mathcal{A}_i}(\mathbb{D})$ of each attribute $\mathcal{A}_i \in \mathbb{C}$.
3. Repeat the following nine steps (steps 4 to 12) until $\mathbb{C} = \emptyset$ or desired number of attributes are selected.
4. Select attribute $\mathcal{A}_i$ from $\mathbb{C}$ as the representative of cluster $\mathbb{V}_i$ that has highest relevance value. In effect, $\mathcal{A}_i \in \mathbb{S}$, $\mathcal{A}_i \in \mathbb{V}_i$, $\mathcal{A}_i \in \bar{\mathbb{V}}_i$, and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_i$.
5. Generate coarse cluster $\mathbb{V}_i$ from the set of existing attributes of $\mathbb{C}$ satisfying the following condition:

$$\mathbb{V}_i = \{\mathcal{A}_j | \Psi(\mathcal{A}_i, \mathcal{A}_j) \geq \delta; \mathcal{A}_j \neq \mathcal{A}_i \in \mathbb{C}\}.$$

6. Initialize $\bar{\mathcal{A}}_i \leftarrow \mathcal{A}_i$.
7. Repeat following four steps (steps 8 to 11) for each attribute $\mathcal{A}_j \in \mathbb{V}_i$.
8. Compute two augmented cluster representatives by averaging $\mathcal{A}_j$ and its complement with the attributes of $\bar{\mathbb{V}}_i$ as follows:

$$\bar{\mathcal{A}}^{+}_{i+j} = \frac{1}{|\bar{\mathbb{V}}_i| + 1} \left\{ \sum_{\mathcal{A}_k \in \bar{\mathbb{V}}_i} \mathcal{A}_k + \mathcal{A}_j \right\}, \tag{14}$$

$$\bar{\mathcal{A}}^{-}_{i+j} = \frac{1}{|\bar{\mathbb{V}}_i| + 1} \left\{ \sum_{\mathcal{A}_k \in \bar{\mathbb{V}}_i} \mathcal{A}_k - \mathcal{A}_j \right\}. \tag{15}$$

9. The augmented cluster representative $\bar{\mathcal{A}}_{i+j}$ after averaging $\mathcal{A}_j$ or its complement with $\bar{\mathbb{V}}_i$ is as follows:

$$\bar{\mathcal{A}}_{i+j} = \begin{cases} \bar{\mathcal{A}}^{+}_{i+j}, & \text{if } \mathrm{R}_{\bar{\mathcal{A}}^{+}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathcal{A}}^{-}_{i+j}}(\mathbb{D}), \\ \bar{\mathcal{A}}^{-}_{i+j}, & \text{otherwise.} \end{cases} \tag{16}$$

10. The augmented cluster representative $\bar{\mathcal{A}}_i$ of cluster $\mathbb{V}_i$ is $\bar{\mathcal{A}}_{i+j}$ if $\mathrm{R}_{\bar{\mathcal{A}}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathcal{A}}_i}(\mathbb{D})$, otherwise $\bar{\mathcal{A}}_i$ remains unchanged.
11. Select attribute $\mathcal{A}_j$ or its complement as a member of the finer cluster $\bar{\mathbb{V}}_i$ of attribute $\mathcal{A}_i$ if $\mathrm{R}_{\bar{\mathcal{A}}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathcal{A}}_i}(\mathbb{D})$.
12. In effect, $\bar{\mathcal{A}}_i \in \bar{\mathbb{S}}$ and $\mathbb{C} = \mathbb{C} \setminus \mathbb{V}_i$.

13. Sort the set of augmented cluster representatives $\mathbb{S} = \{\bar{\mathcal{A}}_i\}$ according to their relevance value $\mathrm{R}_{\bar{\mathcal{A}}_i}(\mathbb{D})$ with respect to the class labels $\mathbb{D}$.
14. Stop.

3.3 Fundamental Property

From the above discussions, the following properties corresponding to each cluster $\mathbb{W}_i$ can be derived:

1. $\Psi(\mathcal{A}_i, \mathcal{A}_j) \geq \delta; \forall \mathcal{A}_j \in \mathbb{W}_i$.
2. $\mathrm{R}_{\mathcal{A}_i}(\mathbb{D}) \geq \mathrm{R}_{\mathcal{A}_j}(\mathbb{D}); \forall \mathcal{A}_j \in \mathbb{W}_i$.
3. $\mathrm{R}_{\bar{\mathcal{A}}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathcal{A}}_i}(\mathbb{D}); \forall \mathcal{A}_j \in \bar{\mathbb{W}}_i$.
4. $\mathrm{R}_{\bar{\mathcal{A}}_{i+j}}(\mathbb{D}) < \mathrm{R}_{\bar{\mathcal{A}}_i}(\mathbb{D}); \forall \mathcal{A}_j \in \mathbb{W}_i \setminus \bar{\mathbb{W}}_i$.
5. $\mathbb{W}_i \cap \mathbb{W}_k = \emptyset, \forall i \neq k$.

The property 1 says that if an attribute $\mathcal{A}_j \in \mathbb{W}_i \Rightarrow \Psi(\mathcal{A}_i, \mathcal{A}_j) \geq \delta$. That is, the supervised similarity between the attribute $\mathcal{A}_j$ of coarse cluster $\mathbb{W}_i$ and the initial cluster representative $\mathcal{A}_i$ is greater than a predefined threshold value δ. The property 2 establishes the fact that if $\mathcal{A}_j \in \mathbb{W}_i \Rightarrow \mathrm{R}_{\mathcal{A}_i}(\mathbb{D}) \geq \mathrm{R}_{\mathcal{A}_j}(\mathbb{D})$, that is, the relevance of the cluster representative $\mathcal{A}_i$ is the maximum among that of all attributes of the cluster $\mathbb{W}_i$. The properties 3 and 4 are of great importance in increasing the relevance of augmented cluster representative with respect to the class labels and reducing the redundancy among the attribute set. The property 3 says that if $\mathcal{A}_j \in \bar{\mathbb{W}}_i \Rightarrow \mathrm{R}_{\bar{\mathcal{A}}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathcal{A}}_i}(\mathbb{D})$. It means an attribute $\mathcal{A}_j$ belongs to the finer cluster $\bar{\mathbb{W}}_i$ if and only if it increases the relevance value of the augmented cluster representative $\bar{\mathcal{A}}_i$. On the other hand, property 4 says that the attributes those belong to only coarse cluster $\mathbb{W}_i$, not to finer cluster $\bar{\mathbb{W}}_i$, are not responsible to increase the relevance of augmented cluster representative. Hence, the set of attributes $\bar{\mathbb{W}}_i$ increases the relevance value of the attribute $\mathcal{A}_i$ as well as reduces the redundancy of the whole set, while the set of attributes $\mathbb{W}_i \setminus \bar{\mathbb{W}}_i$ is only responsible for reducing the redundancy. Finally, property 5 says that if an attribute $\mathcal{A}_i \in \mathbb{W}_i \Rightarrow \mathcal{A}_i \notin \mathbb{W}_k, \forall k \neq i$, that is, the attribute $\mathcal{A}_i$ is contained in $\mathbb{W}_i$ only. Hence, the proposed algorithm generates nonoverlapping attribute clusters.

3.4 Computational Complexity

The computation of the relevance of m attributes is carried out in step 2 of the proposed algorithm, which has $\mathcal{O}(m)$ time complexity. The cluster generation steps, that is steps 4 to 12 are executed c times to generate c clusters and corresponding augmented cluster representatives. There are three loops in the cluster generation steps, which are executed m, m, and m_i times, respectively, where $m_i < m$ represents the cardinality of the cluster $\mathbb{W}_i$. Each iteration of the loops takes only a constant amount of time. Hence, the complexity to generate c clusters using steps 4 to 12 is $\mathcal{O}(c(m + m_i))$. The computing time of $\mathcal{O}(c(m + m_i))$ becomes $\mathcal{O}(cm)$ for any value of m_i. Finally, step 13 performs the sorting of c augmented cluster representatives according to their relevance values, which has a computational complexity of $\mathcal{O}(c^2)$. Hence, the overall time complexity of the proposed supervised clustering algorithm is $\mathcal{O}(m + cm + c^2)$, that is, $\mathcal{O}(cm + c^2)$. However, as the number of desired clusters c is constant and sufficiently small compared to the total number of attributes m, the proposed clustering algorithm has an overall $\mathcal{O}(m)$ time complexity.

4 Experimental Results

The performance of the proposed supervised attribute clustering algorithm is extensively compared with that of some existing supervised and unsupervised gene clustering and gene selection algorithms, namely, ACA (attribute clustering algorithm) [11], MBBC (model-based bayesian clustering) [14], SGCA (supervised gene clustering algorithm) [21], GS (gene shaving) [22], mRMR (minimum redundancy-maximum relevance framework) [25], and the method proposed by Golub et. al [1]. To analyze the performance of different algorithms, the experimentation is done on five microarray gene expression data sets. The major metrics for evaluating the performance of different algorithms are the class separability index [5] and classification accuracy of naive bayes classifier [9], K-nearest neighbor rule [9], and support vector machine [36]. To compute the classification accuracy, the leave-one-out cross-validation is performed on each gene expression data set.

The proposed algorithm is implemented in C language and run in LINUX environment having machine configuration Pentium IV, 3.2 GHz, 1 MB cache, and 1 GB RAM. The kernel-based method is used to approximate probability density functions by combining basis functions [35]. It consists in superposing a Gaussian function to each point of the feature. The final probability density function approximation is obtained by taking the envelope of all the basis functions superposed at each point. The gnu scientific library is used to implement the kernel-based approach.

4.1 Gene Expression Data Sets

In this paper, publicly available three cancer and two arthritis data sets are used. Since binary classification is a typical and fundamental issue in diagnostic and prognostic prediction of cancer and arthritis, different methods are compared using the following five binary-class data sets.

1. *Breast Cancer*. The breast cancer data set contains expression levels of 7,129 genes in 49 breast tumor samples [37]. 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 7,070 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 2,000 genes and 62 samples from two classes [38]: 40 tumor and 22 normal colon tissues.
4. *Rheumatoid Arthritis versus Osteoarthritis*. The RAOA data set consists of gene expression profiles of thirty patients: 21 with RA and 9 with OA [39]. The Cy5-labeled experimental cDNA and the Cy3 labeled common reference sample were pooled and hybridized to the lymphochips containing $\sim 18,000$ cDNA spots representing genes of relevance in immunology [39].
5. *Rheumatoid Arthritis versus Healthy Controls*. The RAHC data set consists of gene expression profiling of peripheral blood cells from 32 patients with RA, 3 patients with probable RA, and 15 age and

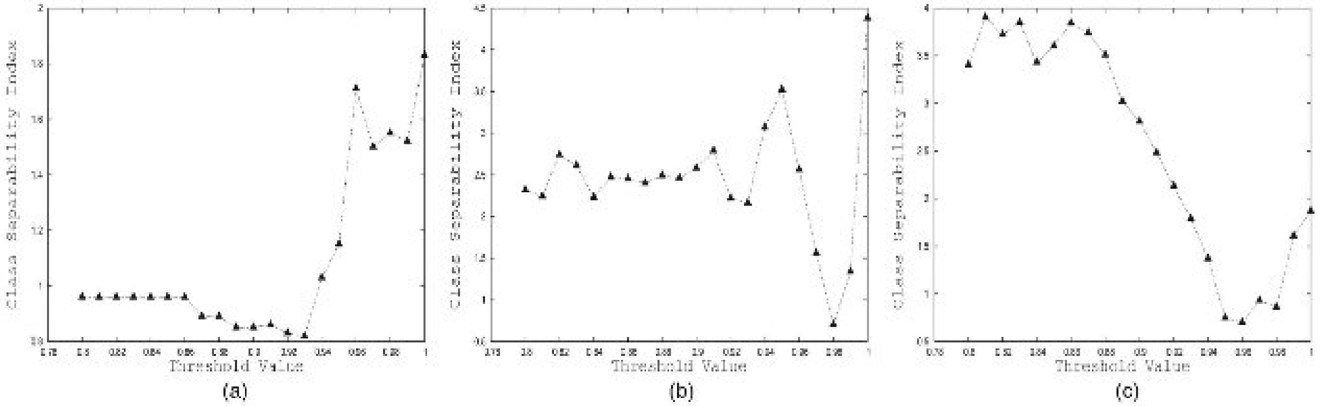


Fig. 2. Variation of class separability index for different values of threshold δ. (a) Colon. (b) RAHC. (c) RAOA.

sex matched healthy controls performed on microarrays with a complexity of ~26 K unique genes (43 K elements) [40].

4.2 Class Prediction Methods

Following three classifiers are used to evaluate the performance of different methods with respect to five microarray data sets.

4.2.1 *Support Vector Machine*

The support vector machine (SVM) [36] is a relatively new and promising classification method. It 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.

4.2.2 *K-Nearest Neighbor Rule*

The K-nearest neighbor rule [9] 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 among its K-nearest neighbors. The value of K, chosen for the K-NN, is the square root of number of samples in training set.

4.2.3 *Naive Bayes Classifier*

The naive bayes classifier [9] 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 $\{w_{1j}, \dots, w_{ij}, \dots, w_{mj}\}$ for the m genes, the posterior probability that x_j belongs to class c is

$$p(c|x_j) \propto \prod_{i=1}^{m} p(w_{ij}|c), \tag{17}$$

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

4.3 Optimum Value of Threshold

The threshold δ in (13) plays an important role to form the initial coarse cluster. It controls the degree of similarity among the attributes of a cluster. In effect, it has a direct influence on the performance of the proposed supervised attribute clustering algorithm. If δ increases, the number of attributes in a cluster decreases, but the similarity among them with respect to sample categories increases. On the other hand, the similarity among the attributes of a cluster decreases with the decrease in the value of δ.

To find out the optimum value of δ, the class separability index [5] is used. The class separability index $\mathcal{S}$ of a data set is defined as $\mathcal{S} = \text{trace}(V_B^{-1} V_W)$, where V_W is the within class scatter matrix and V_B is the between class scatter matrix, defined as follows:

$$V_W = \sum_{j=1}^{C} \pi_j E\{(X - \mu_j)(X - \mu_j)^T | c_j\} = \sum_{j=1}^{C} \pi_j \Sigma_j,$$

$$V_B = \sum_{j=1}^{C} (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^T, \quad \text{and } \bar{\mu} = E\{X\} = \sum_{j=1}^{C} \pi_j \mu_j,$$

where C is the number of classes, π_j is a priori probability that a pattern belongs to class c_j, X is a feature vector, $\bar{\mu}$ is the sample mean vector for the entire data points, μ_j and Σ_j represent the sample mean and covariance matrix of class c_j, respectively, and $E\{\cdot\}$ is the expectation operator. A lower value of $\mathcal{S}$ ensures that classes are well separated by their scatter means.

For five microarray data sets, the value of δ is varied from 0.80 to 1.0 and the class separability index is computed only for best cluster ($c = 1$). Fig. 2 represents the variation of class separability index with respect to different values of threshold δ on colon cancer, RAHC, and RAOA data sets. From the results reported in Fig. 2, it is seen that as the threshold δ increases, the class separability index decreases and attains its minimum value at a particular value of δ. After that the class separability index increases with the increase in the value of δ. Hence, the optimum value of δ for each microarray data set is obtained using the following relation:

$$\delta_{\text{optimum}} = \arg\min_{\delta}\{\mathcal{S}\}. \tag{18}$$

The optimum values of δ obtained using (18) are 0.97, 0.96, 0.93, 0.98, and 0.96 for breast, leukemia, colon, RAHC,

TABLE 1
Performance of Proposed Algorithm on Three Cancer Microarray Data Sets for Different Values of Threshold δ

Data Sets	Value of c	Measure	Different Values of Threshold δ														
			0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
Breast	1	SVM	91.8	89.8	91.8	87.8	91.8	83.7	91.8	93.9	95.9	85.7	91.8	95.9	100	95.9	87.8
		K-NN	98.0	95.9	100	95.9	98.0	95.9	98.0	93.9	98.0	98.0	98.0	95.9	100	95.9	91.8
		NB	100	100	100	100	100	100	100	100	100	100	100	100	100	98.0	91.8
	2	SVM	91.8	91.8	93.9	89.8	91.8	91.8	93.9	100	98.0	98.0	98.0	98.0	100	95.9	87.8
		K-NN	95.9	95.9	93.9	95.9	100	95.9	95.9	100	98.0	95.9	98.0	100	100	95.9	89.8
		NB	100	100	100	100	100	100	100	100	100	100	100	100	100	98.0	89.8
	3	SVM	100	91.8	91.8	91.8	91.8	91.8	93.9	100	98.0	100	100	100	100	95.9	93.9
		K-NN	100	93.9	93.9	95.9	93.9	93.9	95.9	100	100	100	100	100	100	95.9	93.9
		NB	100	100	100	100	100	100	100	100	100	100	100	100	100	98.0	91.8
Leukemia	1	SVM	93.1	91.7	94.4	97.2	97.2	95.8	94.4	95.8	97.2	95.8	97.2	98.6	97.2	98.6	90.3
		K-NN	98.6	98.6	98.6	97.2	98.6	95.8	95.8	98.6	100	100	97.2	100	98.6	98.6	93.1
		NB	100	100	100	100	100	100	100	100	100	100	100	100	100	100	94.4
	2	SVM	91.7	97.2	94.4	98.6	98.6	97.2	95.8	100	98.6	98.6	98.6	100	100	100	94.4
		K-NN	97.2	97.2	98.6	98.6	98.6	97.2	95.8	100	100	98.6	98.6	100	100	100	94.4
		NB	100	100	100	100	100	100	100	100	100	100	100	100	100	100	98.6
	3	SVM	93.1	97.2	93.1	98.6	98.6	100	100	100	100	98.6	100	100	100	100	93.1
		K-NN	97.2	97.2	100	98.6	98.6	98.6	100	100	98.6	98.6	100	100	100	100	94.4
		NB	100	100	100	100	98.6	100	100	100	100	100	100	100	100	100	98.6
Colon	1	SVM	96.8	96.8	100	100	100	96.8	98.4	96.8	100	98.4	88.7	98.4	100	98.4	90.3
		K-NN	98.4	98.4	100	100	96.8	96.8	98.4	98.4	100	98.4	95.2	100	96.8	96.8	90.3
		NB	100	100	100	100	100	100	100	100	100	100	100	100	100	98.4	91.9
	2	SVM	98.4	98.4	100	100	98.4	95.2	98.4	100	100	100	98.4	100	98.4	98.4	90.3
		K-NN	98.4	98.4	98.4	98.4	100	95.2	98.4	100	100	100	100	100	96.8	98.4	90.3
		NB	100	100	100	100	100	100	100	100	100	100	100	100	98.4	98.4	93.6
	3	SVM	100	100	100	100	100	100	98.4	100	100	100	100	100	96.8	96.8	88.7
		K-NN	100	100	100	100	100	100	98.4	100	100	100	100	100	96.8	98.4	91.9
		NB	100	100	100	100	100	100	100	100	100	100	100	100	96.8	96.8	91.9

TABLE 2
Performance of Proposed Algorithm on Two Arthritis Microarray Data Sets for Different Values of Threshold δ

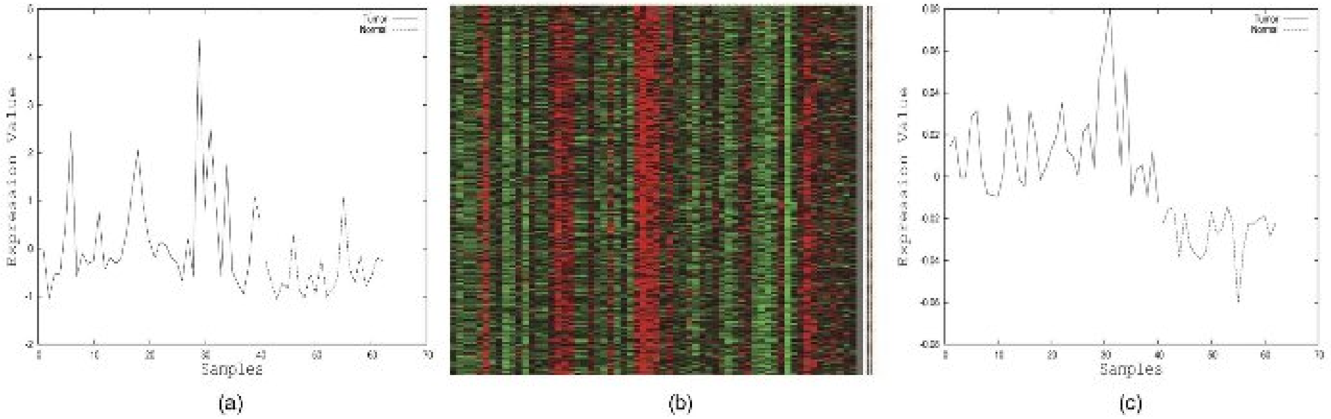
Data Sets	Value of c	Measure	Different Values of Threshold δ														
			0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
RAHC	1	SVM	82.0	78.0	82.0	80.0	78.0	76.0	80.0	82.0	90.0	82.0	70.0	84.0	86.0	100	94.0
		K-NN	96.0	94.0	92.0	98.0	94.0	92.0	96.0	96.0	94.0	92.0	90.0	98.0	98.0	100	94.0
		NB	100	100	100	100	100	100	100	100	100	100	100	100	100	100	94.0
	2	SVM	78.0	98.0	78.0	78.0	78.0	98.0	78.0	98.0	100	88.0	84.0	90.0	100	100	98.0
		K-NN	96.0	98.0	94.0	94.0	94.0	98.0	92.0	98.0	100	94.0	92.0	98.0	100	100	98.0
		NB	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96.0
	3	SVM	98.0	98.0	98.0	98.0	100	98.0	98.0	98.0	100	100	86.0	100	100	100	98.0
		K-NN	96.0	98.0	98.0	96.0	96.0	96.0	100	98.0	100	100	92.0	100	100	100	100
		NB	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96.0
RAOA	1	SVM	73.3	73.3	73.3	76.7	80.0	80.0	76.7	80.0	76.7	76.7	96.7	100	70.0	100	93.3
		K-NN	86.7	96.7	100	93.3	93.3	100	96.7	90.0	96.7	100	96.7	100	83.3	100	96.7
		NB	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96.7
	2	SVM	73.3	76.7	73.3	76.7	80.0	80.0	76.7	80.0	80.0	80.0	96.7	100	96.7	100	93.3
		K-NN	96.7	96.7	93.3	93.3	93.3	90.0	100	86.7	90.0	100	100	100	100	100	93.3
		NB	100	100	100	100	100	100	100	100	100	100	100	100	100	100	96.7
	3	SVM	70.0	70.0	66.7	73.3	76.7	76.7	53.3	83.3	80.0	73.3	96.7	100	96.7	100	96.7
		K-NN	96.7	96.7	93.3	90.0	93.3	90.0	90.0	90.0	100	93.3	100	100	100	100	93.3
		NB	100	100	100	100	100	100	100	100	100	100	100	100	100	100	90.0

and RAOA data sets, respectively. Finally, Tables 1 and 2 present the performance of the proposed clustering algorithm for different values of δ. The results and subsequent discussions are presented with respect to the classification accuracy of the SVM, K-NN rule, and NB classifier. The results are reported for three best clusters ($c = 3$) obtained using the proposed attribute clustering method. From the results reported in Tables 1 and 2, it is also seen that the proposed supervised attribute clustering algorithm achieves its best performance at $\delta = \delta_{\text{optimum}}$, irrespective of the classifiers used. However, the performance of the proposed method at $\delta = 0.98$ is same as that at δ_{optimum} for RAOA data set with respect to the classification accuracy of three classifiers.

4.4 Qualitative Analysis of Supervised Clusters

For three cancer and two arthritis data sets, the best clusters generated by the proposed algorithm are analyzed using the Eisen plot. In Eisen plot [41], the expression value of a gene in a particular sample is represented by coloring the corresponding cell of the data matrix with a color similar to the original color of its spot on the microarray. The shades of red color represent higher expression level, the shades of green color represent lower expression level and the colors toward black represent absence of differential expression values.

In Figs. 3a, 3b, 3c and Figs. 4a, 4b, 4c, the results of best cluster obtained using the proposed clustering algorithm are reported for colon cancer and RAHC data sets considering the values of δ as 0.93 and 0.98, respectively. Figs. 3a and 4a show



(a) (b) (c)

Fig. 3. Results obtained using proposed algorithm for colon cancer data set considering $\delta = 0.93$. (a) Initial expression value. (b) Eisen plot. (c) Augmented expression value.

the expression values of the actual genes or attributes of the best cluster over the samples for two data sets. Figs. 3b and 4b represent the Eisen plot of corresponding finer cluster with actual gene expression values, while Figs. 3c and 4c show the expression values of the augmented cluster representatives of the best cluster for two data sets. In Fig. 5, the expression values of the actual and augmented cluster representatives of the best cluster are presented for breast, leukemia, and RAOA data sets considering δ as 0.97, 0.96, and 0.96, respectively. All the results reported in Figs. 3, 4, and 5 establish the fact that the proposed supervised attribute clustering algorithm can identify groups of coregulated genes with strong association to the sample categories or class labels.

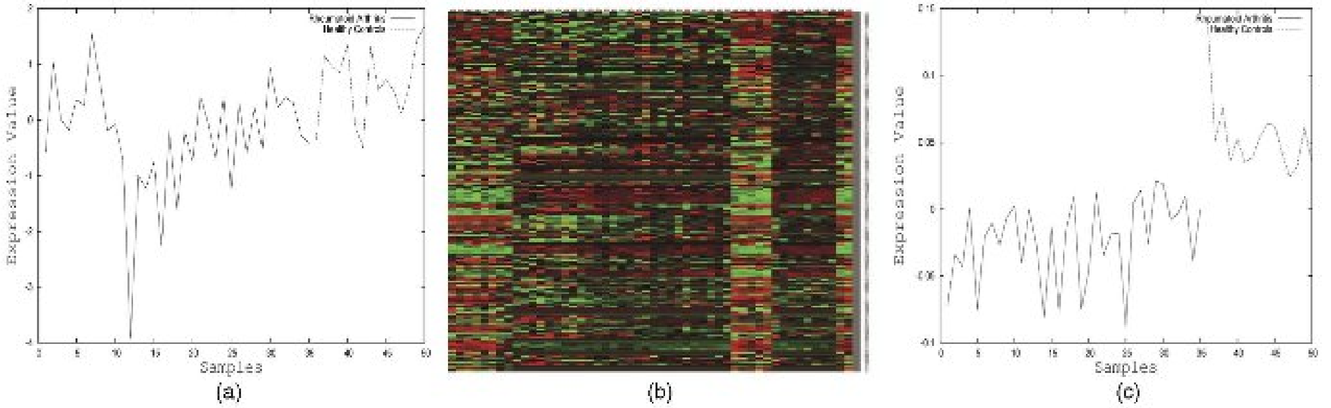
4.5 Importance of Supervised Similarity Measure

The supervised similarity measure based on mutual information, defined in (8), takes into account the information of sample categories or class labels while computing the similarity between two genes. It also incorporates the unsupervised similarity measure among genes. On the other hand, mutual information-based conventional similarity measure of (6) does not consider the class labels or sample categories.

In order to establish the importance of supervised similarity measure over existing conventional unsupervised similarity measure, the extensive experimentation is carried out on three cancer and two arthritis data sets. Finally, the best results obtained using unsupervised similarity measure are compared with that of proposed supervised measure in Table 3 with respect to the class separability index and classification accuracy of the SVM, K-NN rule, and NB classifier. From all the results reported in Table 3, it is seen that the performance of the proposed supervised similarity measure is better compared to that of the unsupervised measure for all microarray data sets. That is, the proposed supervised similarity measure can identify functional groups of genes present in the microarray, while the unsupervised similarity fails to reveal that. However, the unsupervised similarity measure performs better than supervised similarity with respect to the class separability index for $c = 2$ in case of leukemia and RAHC data, and for $c = 3$ in case of leukemia, RAHC, and RAOA data.

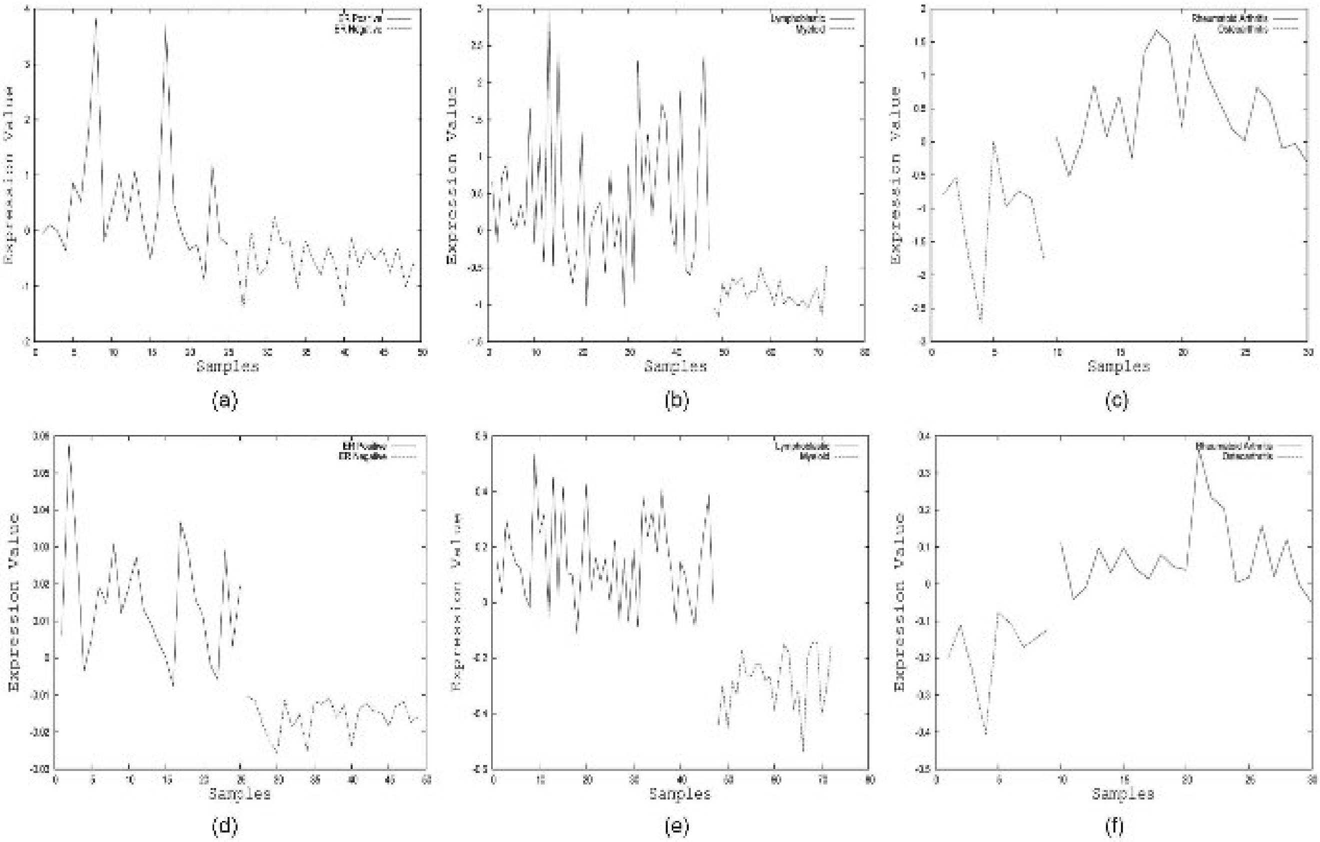
4.6 Importance of Augmented Genes

Each coarse cluster represents the set of genes or attributes those have the supervised similarity values with the initial cluster representative greater than a predefined threshold value δ. In fact, the relevance of the initial cluster representative is greater than that of other genes of that cluster. After forming the initial coarse cluster, the cluster representative is refined incrementally in the proposed algorithm. By searching among the genes of coarse cluster, the current cluster representative is merged and averaged with one single gene



(a) (b) (c)

Fig. 4. Results obtained using proposed algorithm for RAHC data set considering $\delta = 0.98$. (a) Initial expression value. (b) Eisen plot. (c) Augmented expression value.



(a) (b) (c)

(d) (e) (f)

Fig. 5. Results obtained for breast cancer, leukemia, and RAOA data considering $\delta =$ 0.97, 0.96, and 0.96, respectively. (a) Initial representative of breast. (b) Initial representative of leukemia. (c) Initial representative of RAOA. (d) Augmented representative of breast. (e) Augmented representative of leukemia. (f) Augmented representative of RAOA.

such that the augmented cluster representative increases the relevance value. The merging process is repeated until the relevance value can no longer be improved.

In order to establish the importance of augmented cluster representative over initial cluster representative, that is, actual gene, extensive experiments are carried out on five microarray data sets. Table 4 reports the comparative performance of actual and augmented genes of different clusters. Results are reported for $c = 3$ considering supervised similarity measure. The performance of actual and augmented genes is compared with respect to the class separability index and classification accuracy of the SVM, K-NN rule, and NB classifier. All the results reported in Table 4 establish the fact that the proposed supervised attribute clustering algorithm performs significantly better in case of augmented gene than the actual gene. Only in case of leukemia data for $c = 2$ and 3, and RAOA data for $c = 3$, the actual gene performs better than augmented one with respect to the class separability index.

4.7 Comparison between Coarse and Finer Clusters

In the proposed attribute clustering algorithm, the augmented cluster representative is computed by averaging the genes of finer cluster, rather than all genes of corresponding coarse cluster. That is, instead of averaging all genes of

TABLE 3
Comparative Performance Analysis of Supervised and Unsupervised Similarity Measures for Different Data Sets

Value of c	Measure	Breast		Leukemia		Colon		RAHC		RAOA	
		proposed	existing	proposed	existing	proposed	existing	proposed	existing	proposed	existing
1	SVM	100	89.8	98.6	90.3	100	88.7	100	94.0	100	96.7
	K-NN	100	87.8	100	93.1	100	87.1	100	100	100	96.7
	NB	100	89.8	100	94.4	100	88.7	100	100	100	100
	CS	0.53	1.09	0.49	1.00	0.82	1.33	0.70	0.72	0.70	1.14
2	SVM	100	87.8	100	94.4	100	83.9	100	100	100	93.3
	K-NN	100	85.7	100	94.4	100	85.5	100	100	100	93.3
	NB	100	89.8	100	98.6	100	88.7	100	100	100	100
	CS	0.50	1.65	0.98	0.77	0.94	1.56	0.87	0.66	0.99	1.09
3	SVM	100	91.8	100	93.1	100	87.1	100	98.0	100	93.3
	K-NN	100	91.8	100	94.4	100	87.1	100	98.0	100	96.7
	NB	100	95.9	100	98.6	100	88.7	100	98.0	100	96.7
	CS	0.61	1.51	1.12	0.80	0.97	1.84	0.88	0.76	1.96	1.07

TABLE 4
Comparative Performance Analysis of Augmented and Actual Genes for Different Data Sets

Value of c	Measure	Breast		Leukemia		Colon		RAHC		RAOA	
		augmented	actual	augmented	actual	augmented	actual	augmented	actual	augmented	actual
1	SVM	100	85.7	98.6	90.3	100	83.9	100	88.0	100	93.3
	K-NN	100	89.8	100	93.1	100	83.9	100	88.0	100	90.0
	NB	100	89.8	100	94.4	100	83.9	100	68.0	100	93.3
	CS	0.53	1.16	0.49	1.00	0.82	1.83	0.70	4.38	0.70	0.87
2	SVM	100	85.7	100	94.4	100	83.9	100	72.0	100	90.0
	K-NN	100	91.8	100	94.4	100	83.9	100	84.0	100	86.7
	NB	100	89.8	100	98.6	100	88.7	100	88.0	100	86.7
	CS	0.50	1.56	0.98	0.77	0.94	1.86	0.87	7.93	0.99	1.07
3	SVM	100	87.8	100	93.1	100	80.6	100	84.0	100	93.3
	K-NN	100	89.8	100	94.4	100	85.5	100	86.0	100	90.0
	NB	100	91.8	100	98.6	100	83.9	100	94.0	100	83.3
	CS	0.61	1.97	1.12	0.80	0.97	2.31	0.88	4.03	1.96	1.27

coarse cluster, the augmented gene is computed by considering a subset of genes of coarse cluster, which is termed as the finer cluster, those increase the relevance value of initial cluster representative.

Table 5 presents the comparative performance of means computed from coarse cluster and that from finer cluster. The comparison is reported for $c = 3$ with respect to the class separability index and classification accuracy of the SVM, K-NN rule, and NB classifier. The results reported in Table 5 establish the fact that the augmented cluster representative obtained from finer cluster performs significantly better than that of coarse cluster, irrespective of the data sets and quantitative indices used. The attributes those present in the coarse cluster, but not in the corresponding finer cluster, are not responsible to increase the relevance value with respect to the class labels or response variables. Also, they degrade the quality of solution. Hence, the augmented cluster representatives should be computed by considering only genes of finer clusters, not all genes of coarse clusters.

4.8 Comparative Performance Analysis

Finally, Table 6 compares the best performance of the proposed algorithm with that of some existing algorithms such as ACA [11], MBBC [14], SGCA [21], GS [22], and mRMR [25]. The results are presented based on the best classification accuracy of the SVM, K-NN rule, and NB classifier for five microarray data sets. The values of δ are considered as 0.97, 0.96, 0.93, 0.98, and 0.96 for breast cancer, leukemia, colon cancer, RAHC, and RAOA data sets, respectively. From the results reported in Table 6, it is seen that the proposed supervised gene clustering algorithm generates a set of clusters having highest classification accuracy of the SVM, K-NN rule, and NB classifier, and lowest class separability index values for all the cases. The better performance of the proposed clustering algorithm is achieved due to the fact that it can identify functional groups of genes present in the microarray data sets more accurately than the existing algorithms. However, with respect to the class separability index, mRMR [25] for $c = 3$ and GS [22] for $c = 1$ perform better than the proposed method in case of leukemia data and for RAOA data at $c = 3$, both ACA [11] and SGCA [21] attain lower class separability index values than the proposed algorithm. In this regard, it should be noted that the method proposed by Golub et. al [1] achieves maximum accuracy of 98.0, 98.6, 91.9, 98.0, and 96.7 percent for breast, leukemia, colon, RAHC, and RAOA data sets, respectively.

4.9 Biological Significance

To interpret the biological significance of the generated clusters, the Gene Ontology (GO) Term Finder is used [42]. 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

TABLE 5
Comparative Performance Analysis of Means of Coarse and Finer Clusters for Different Data Sets

Value of c	Measure	Breast		Leukemia		Colon		RAHC		RAOA	
		finer	coarse	finer	coarse	finer	coarse	finer	coarse	finer	coarse
1	SVM	100	63.3	98.6	66.7	100	64.5	100	70.0	100	70.0
	K-NN	100	69.4	100	66.7	100	62.9	100	64.0	100	66.7
	NB	100	63.3	100	73.6	100	61.3	100	60.0	100	80.0
	CS	0.53	5.38	0.49	7.69	0.82	68.47	0.70	11.14	0.70	119.53
2	SVM	100	59.2	100	72.2	100	64.5	100	70.0	100	70.0
	K-NN	100	55.1	100	68.1	100	61.3	100	58.0	100	66.7
	NB	100	59.2	100	68.1	100	61.3	100	66.0	100	80.0
	CS	0.50	7.63	0.98	13.66	0.94	70.62	0.87	12.74	0.99	118.23
3	SVM	100	59.2	100	73.6	100	64.5	100	82.0	100	70.0
	K-NN	100	55.1	100	65.3	100	61.3	100	82.0	100	66.7
	NB	100	67.4	100	75.0	100	61.3	100	62.0	100	80.0
	CS	0.61	8.85	1.12	20.95	0.97	72.95	0.88	12.62	1.96	138.51

TABLE 6
Comparative Performance Analysis of Different Methods for Five Microarray Data Sets

Data Sets	Methods / Algorithms	$c = 1$				$c = 2$				$c = 3$			
		SVM	K-NN	NB	CS	SVM	K-NN	NB	CS	SVM	K-NN	NB	CS
Breast	Proposed	100	100	100	0.53	100	100	100	0.50	100	100	100	0.61
	ACA	81.6	81.6	81.6	2.07	81.6	83.7	81.6	2.92	89.8	83.7	83.7	1.03
	MBBC	79.6	79.6	85.7	1.04	79.6	81.6	85.7	1.18	81.6	81.6	89.8	0.94
	SGCA	100	100	100	1.74	100	100	100	1.29	100	100	100	1.83
	GS	75.5	79.6	79.6	1.68	85.7	85.7	83.7	2.60	89.8	87.8	85.7	3.75
	mRMR	85.7	89.8	89.8	1.16	81.6	89.8	95.9	1.70	93.9	95.9	100	1.54
Leukemia	Proposed	98.6	100	100	0.49	100	100	100	0.98	100	100	100	1.12
	ACA	82.4	82.4	88.2	1.69	88.2	82.4	88.2	1.19	88.2	91.2	88.2	3.25
	MBBC	88.2	88.2	90.3	0.94	94.4	91.7	93.1	1.89	94.4	93.1	95.8	1.63
	SGCA	93.1	94.4	94.4	1.16	94.4	94.4	94.4	2.01	94.4	95.8	94.4	1.76
	GS	97.2	94.4	91.7	0.43	97.2	97.2	95.8	1.67	100	100	94.4	1.90
	mRMR	90.3	93.1	94.4	1.00	94.4	94.4	98.6	1.46	94.4	95.8	100	1.08
Colon	Proposed	100	100	100	0.82	100	100	100	0.94	100	100	100	0.97
	ACA	72.6	77.4	64.5	3.08	72.6	83.9	75.8	1.46	77.4	83.9	64.5	2.59
	MBBC	64.5	64.5	72.6	1.68	75.8	72.6	72.6	1.69	75.8	75.8	82.3	3.05
	SGCA	72.6	72.6	75.8	5.10	75.8	77.4	77.4	3.80	77.4	77.4	77.4	4.25
	GS	83.9	82.3	82.3	1.41	82.3	83.9	79.0	2.70	87.1	87.1	85.5	4.10
	mRMR	83.9	83.9	83.9	1.83	83.9	83.9	83.9	2.51	75.8	83.9	83.9	3.89
RAHC	Proposed	100	100	100	0.70	100	100	100	0.87	100	100	100	0.88
	ACA	90.0	88.0	88.0	2.79	90.0	96.0	92.0	4.81	92.0	92.0	92.0	3.02
	MBBC	86.0	84.0	84.0	1.15	84.0	88.0	90.0	2.09	90.0	92.0	90.0	1.77
	SGCA	92.0	92.0	92.0	1.76	90.0	96.0	92.0	3.08	90.0	96.0	92.0	2.17
	GS	64.0	68.0	62.0	5.26	84.0	82.0	72.0	8.78	88.0	88.0	66.0	13.23
	mRMR	88.0	88.0	68.0	4.38	84.0	90.0	96.0	3.57	92.0	90.0	98.0	3.35
RAOA	Proposed	100	100	100	0.70	100	100	100	0.99	100	100	100	1.96
	ACA	86.7	83.3	83.3	1.90	86.7	83.3	83.3	2.11	86.7	86.7	86.7	1.54
	MBBC	86.7	86.7	83.3	1.91	86.7	86.7	90.0	2.06	86.7	86.7	86.7	3.88
	SGCA	93.3	90.0	90.0	1.71	93.3	93.3	96.7	3.04	93.3	93.3	96.7	1.67
	GS	80.0	73.3	83.3	1.46	93.3	96.7	80.0	2.94	86.7	93.3	83.3	4.61
	mRMR	93.3	90.0	93.3	0.87	96.7	96.7	90.0	1.26	96.7	100	90.0	2.01

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 [42]. 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. The FDR is particularly useful in the analysis of high-throughput data such as microarray gene expression.

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 both p-value and FDR (percent) 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 7 presents the significant shared GO terms for the BP, along with the p-values and FDR for the BP, MF, and CC on different data sets. The results corresponding to the best clusters of some existing algorithms such as GS [22] and SGCA [21] are also provided on same data sets for the sake of comparison. The "$*$" in Table 7 represents that no significant shared term is found considering p-value cutoff as 0.05. From the results reported in Table 7, it is seen that the best cluster generated by the proposed algorithm can be

TABLE 7
Significant Shared GO Terms for Genes in Best Clusters Obtained by Different Methods

Data Sets	Methods / Algorithms	Biological Process			Molecular Function		Cellular Component	
		Gene Ontology Term	p-value	FDR	p-value	FDR	p-value	FDR
Breast	Proposed	Positive regulation of biological process	5.7E-027	0	6.9E-024	0	2.1E-020	0
	GS	*	*	*	2.7E-03	4	*	*
	SGCA	*	*	*	*	*	*	*
Leukemia	Proposed	Multicellular organismal development	2.1E-02	10	2.4E-03	2	2.1E-03	0
	GS	*	*	*	2.3E-03	2	1.9E-03	4
	SGCA	*	*	*	4.8E-03	0	*	*
Colon	Proposed	Cellular process	1.8E-012	0	9.0E-016	0	2.0E-09	0
	GS	Regulation of system process	1.2E-02	32	4.5E-03	0	1.9E-03	0
	SGCA	Blood vessel development	1.7E-02	20	*	*	*	*
RAOA	Proposed	Immune system process	8.3E-07	0	2.5E-03	2	*	*
	GS	Immune system process	2.9E-016	0	9.9E-04	0	7.9E-04	0
	SGCA	Interspecies interaction between organisms	7.4E-03	18	3.0E-03	2	*	*

assigned to the GO biological processes with high reliability in terms of p-value and FDR. That is, the proposed algorithm describes accurately the known classification, the one given by the GO, and thus reliable for extracting new biological insights.

5 Conclusion

The main contribution of this paper is threefold, namely,

1. Defining a new quantitative measure, based on mutual information, to calculate the similarity between two genes, which incorporates the information of sample categories or class labels.
2. Development of a new supervised attribute clustering algorithm to find coregulated clusters of genes whose collective expression is strongly associated with the sample categories.
3. Comparing the performance of the proposed method and some existing methods using the class separability index and predictive accuracy of support vector machine, K-nearest neighbor rule, and naive bayes classifier.

For five microarray data, significantly better results are found for the proposed method compared to existing methods, irrespective of the classifiers used. All the results reported in this paper demonstrate the feasibility and effectiveness of the proposed method. It is capable of identifying coregulated clusters of genes whose average expression is strongly associated with the sample categories. The identified gene clusters may contribute to revealing underlying class structures, providing a useful tool for the exploratory analysis of biological data.

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. Statistical Physics*, vol. 110, nos. 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] 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.

[5] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Prentice Hall, 1982.

[6] D. Koller and M. Sahami, "Toward Optimal Feature Selection," *Proc. Int'l Conf. Machine Learning*, pp. 284-292. 1996.

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

[8] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*. Prentice Hall, 1988.

[9] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1999.

[10] D. Jiang, C. Tang, and A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey," *IEEE Trans. Knowledge and Data Eng.*, vol. 16, no. 11, pp. 1370-1386, Nov. 2004.

[11] W.-H. Au, K.C.C. Chan, A.K.C. Wong, and Y. Wang, "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 2, no. 2, pp. 83-101, Apr.-June 2005.

[12] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G.C. Tseng, "Evaluation and Comparison of Gene Clustering Methods in Microarray Analysis," *Bioinformatics*, vol. 22, no. 19, pp. 2405-2412, 2006.

[13] M. Medvedovic and S. Sivaganesan, "Bayesian Infinite Mixture Model Based Clustering of Gene Expression Profiles," *Bioinformatics*, vol. 18, no. 9, pp. 1194-1206, 2002.

[14] Y. Joo, J.G. Booth, Y. Namkoong, and G. Casella, "Model-Based Bayesian Clustering (MBBC)," *Bioinformatics*, vol. 24, no. 6, pp. 874-875, 2008.

[15] 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.

[16] W. Haiying, Z. Huiru, and A. Francisco, "Poisson-Based Self-Organizing Feature Maps and Hierarchical Clustering for Serial Analysis of Gene Expression Data," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 4, no. 2, pp. 163-175, Apr.-June 2007.

[17] L.J. Heyer, S. Kruglyak, and S. Yooseph, "Exploring Expression Data: Identification and Analysis of Coexpressed Genes," *Genome Research*, vol. 9, no. 11, pp. 1106-1115, 1999.

[18] 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. Nat'l Academy of Science USA*, vol. 96, no. 6, pp. 2907-2912, 1999.

[19] K.Y. Yeung and W.L. Ruzzo, "Principal Component Analysis for Clustering Gene Expression Data," *Bioinformatics*, vol. 17, no. 9, pp. 763-774, 2001.

[20] G.J. McLachlan, K.-A. Do, and C. Ambroise, *Analyzing Microarray Gene Expression Data*. Wiley-Interscience, 2004.

[21] M. Dettling and P. Buhlmann, "Supervised Clustering of Genes," *Genome Biology*, vol. 3, no. 12, pp. 0069.1-0069.15, 2002.

[22] T. Hastie, R. Tibshirani, M.B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W.C. Chan, D. Botstein, and P. Brown, "'Gene Shaving' as a Method for Identifying Distinct Sets of Genes with Similar Expression Patterns," *Genome Biology*, vol. 1, no. 2, pp. 1-21, 2000.

[23] T. Hastie, R. Tibshirani, D. Botstein, and P. Brown, "Supervised Harvesting of Expression Trees," *Genome Biology*, vol. 1, pp. 1-12, 2001.

[24] D. Nguyen and D. Rocke, "Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data," *Bioinformatics*, vol. 18, pp. 39-50, 2002.

[25] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *J. Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185-205, 2005.

[26] H. Peng, F. Long, and C. Ding, "Feature Selection Based 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. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537-550, July 1994.

[28] D. Huang and T.W.S. Chow, "Effective Feature Selection Scheme Using Mutual Information," *Neurocomputing*, vol. 63, pp. 325-343, 2004.

[29] X. Liu, A. Krishnan, and A. Mondry, "An Entropy Based Gene Selection Method for Cancer Classification Using Microarray Data," *BMC Bioinformatics*, vol. 6, no. 76, pp. 1-14, 2005.

[30] I. Dhillon, S. Mallela, and R. Kumar, "Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification," *J. Machine Learning Research*, vol. 3, pp. 1265-1287, 2003.

[31] R. Jornsten and B. Yu, "Simultaneous Gene Clustering and Subset Selection for Sample Classification via MDL," *Bioinformatics*, vol. 19, no. 9, pp. 1100-1109, 2003.

[32] J. Li, H. Su, H. Chen, and B.W. Futscher, "Optimal Search-Based Gene Subset Selection for Gene Array Cancer Classification," *IEEE Trans. Information Technology in Biomedicine*, vol. 11, no. 4, pp. 398-405, July 2007.

[33] P. Maji, "f-Information Measures for Efficient Selection of Discriminative Genes from Microarray Data," *IEEE Trans. Biomedical Eng.*, vol. 56, no. 4, pp. 1063-1069, Apr. 2009.

[34] C. Shannon and W. Weaver, *The Math. Theory of Communication* Univ. Illinois Press, 1964.

[35] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

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

[37] 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. Nat'l Academy of Science USA,* vol. 98, no. 20, pp. 11462-11467, 2001.

[38] 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. Nat'l Academy of Science USA,* vol. 96, no. 12, pp. 6745-6750, 1999.

[39] T.C.T.M. van der Pouw Kraan, F.A. van Gaalen, P.V. Kasperkovitz, N.L. Verbeet, T.J.M. Smeets, M.C. Kraan, M. Fero, P.-P. Tak, T.W.J. Huizinga, E. Pieterman, F.C. Breedveld, A.A. Alizadeh, and C.L. Verweij, "Rheumatoid Arthritis is a Heterogeneous Disease: Evidence for Differences in the Activation of the STAT-1 Pathway between Rheumatoid Tissues," *Arthritis and Rheumatism,* vol. 48, no. 8, pp. 2132-2145, 2003.

[40] T.C.T.M. van der Pouw Kraan, C.A. Wijbrandts, L.G.M. van Baarsen, A.E. Voskuyl, F. Rustenburg, J.M. Baggen, S.M. Ibrahim, M. Fero, B.A.C. Dijkmans, P.P. Tak, and C.L. Verweij, "Rheumatoid Arthritis Subtypes Identified by Genomic Profiling of Peripheral Blood Cells: Assignment of a Type I Interferon Signature in a Subpopulation of Pateints," *Annals of the Rheumatic Diseases,* vol. 66, pp. 1008-1014, 2007.

[41] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc. Nat'l Academy of Sciences USA,* vol. 95, no. 25, pp. 14863-14868, 1998.

[42] 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 BSc degree in physics, the MSc degree in electronics science, and the PhD 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, India. His research interests include pattern recognition, computational biology and bioinformatics, medical image processing, cellular automata, soft computing, and so forth. He has published around 60 papers in international journals and conferences. He is also a reviewer of many international journals. He has received the 2006 Best Paper Award of the International Conference on Visual Information Engineering from The Institution of Engineering and Technology, United Kingdom, 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 Associate of the Indian Academy of Sciences, India.