Fuzzy–Rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data

Pradipta Maji

Abstract—**One of the major tasks with gene expression data is to find groups of coregulated genes whose collective expression is strongly associated with sample categories. In this regard, a new clustering algorithm, termed as fuzzy–rough supervised attribute clustering (FRSAC), is proposed to find such groups of genes. The proposed algorithm is based on the theory of fuzzy–rough sets, which directly incorporates the information of sample categories into the gene clustering process. A new quantitative measure is introduced based on fuzzy–rough sets that incorporates the information of sample categories to measure the similarity among genes. The proposed algorithm is based on measuring the similarity between genes using the new quantitative measure, whereby redundancy among the genes is removed. The clusters are refined incrementally based on sample categories. The effectiveness of the proposed FRSAC algorithm, along with a comparison with existing supervised and unsupervised gene selection and clustering algorithms, is demonstrated on six cancer and two arthritis data sets based on the class separability index and predictive accuracy of the naive Bayes' classifier, the K-nearest neighbor rule, and the support vector machine.**

Index Terms—**Attribute clustering, classification, gene selection, microarray analysis, rough sets.**

I. Introduction

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 [1]. An important application of microarray gene expression data in functional genomics is to classify samples according to their gene expression profiles. 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. Hence, identifying a reduced set of most relevant genes is the goal of gene selection. As this is a feature selection problem, a clustering method can be used, which partitions the given gene set into subgroups, each of which should be as homogeneous as possible [2].

Cluster analysis is a technique for finding natural groups present in the data. It divides a given data set into a set of clusters in such a way that two objects from the same cluster are as similar as possible, and the objects from different clusters are as dissimilar as possible [2], [3]. When applied to gene expression data analysis, clustering algorithms can be applied on both gene and sample dimensions [4], [5]. The conventional attribute-clustering methods 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 [5]. 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.

The conventional gene clustering methods allow genes with similar expression patterns, that is, coexpressed genes, to be identified [4]. Different unsupervised clustering techniques such as hierarchical clustering [6], the k-means algorithm [7], the self-organizing map [8], and principal component analysis [9] have been widely applied to find groups of coregulated genes on microarray data. 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 [6]. In contrast, the self-organizing map [8] and the k-means algorithm [7] 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 class labels [10].

To reveal groups of coregulated genes with strong association to the sample categories, different supervised attribute-clustering algorithms have been proposed recently [10]–[12]. The supervised attribute clustering is defined as the grouping of attributes, controlled by the values of attributes as well as the supervised information of sample categories [10]. Previous work in this field encompasses tree harvesting [11], a two-step method that 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.

An interesting supervised clustering approach that directly incorporates the response variables in the grouping process is the partial least squares procedure [12], which, in a supervised manner, constructs weighted linear combinations of genes that have maximal covariance with the outcome. However, it has the

Manuscript received September 24, 2009; revised February 22, 2010; accepted May 3, 2010. Date of publication June 10, 2010; date of current version January 14, 2011. This paper was recommended by Associate Editor A. F. Gomez Skarmeta.

The author is with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India (e-mail: pmaji@isical.ac.in).

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 [10]. The supervised attribute clustering algorithm proposed in [10] 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 [10], 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 the noise or the outlier of the data set.

One of the main problems in gene expression data analysis is uncertainty. Some of the sources of this uncertainty include incompleteness and vagueness in class definitions. In this background, the possibility concept introduced by fuzzy sets [13] and rough sets [14] has gained popularity in modeling and propagating uncertainty. Both fuzzy sets and rough sets provide a mathematical framework to capture uncertainties associated with the data [15]. The generalized theories of rough–fuzzy sets and fuzzy–rough sets have been applied successfully to feature selection of real valued data [16], [17], rough–fuzzy clustering [18], and so forth. To cluster coexpressed genes from microarray data, different fuzzy [19], [20] and rough–fuzzy [18] clustering algorithms can be used. However, these algorithms are unsupervised in nature, as the genes are clustered without using any information of class labels. Recently, the fuzzy–rough set-based feature selection algorithm has been proposed in [21] to select a set of relevant and nonredundant genes from microarray data.

In this paper, a new supervised gene clustering algorithm, termed as fuzzy–rough supervised attribute clustering (FRSAC), is proposed based on the theory of fuzzy–rough sets. It finds coregulated clusters of genes whose collective expression is strongly associated with the sample categories. A new quantitative measure, based on fuzzy–rough sets, is introduced to compute the similarity between genes. This measure incorporates the information of sample categories or class labels while measuring the similarity between genes. The proposed FRSAC 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 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 FRSAC algorithm yields clusters typically made up of a few genes, whose coherent average expression levels allow perfect discrimination of sample categories. The performance of the proposed gene clustering algorithm, along with a comparison with existing algorithms, is studied using the class separability (CS) index and predictive accuracy of naive Bayes' (NB) classifier, the K-nearest neighbor (K-NN) rule, and the support vector machine (SVM) on six cancer and two arthritis data sets.

The structure of the rest of this paper is as follows: Section II briefly introduces rough sets and fuzzy–rough sets. The proposed fuzzy–rough set-based supervised gene clustering algorithm is presented in Section III. A few case studies and a comparison with existing algorithms are presented in Section IV. Concluding remarks are given in Section V.

II. Rough Sets and Fuzzy–Rough Sets

Let $\langle \mathbb{U}, \mathbb{A} \rangle$ represent an approximation space or an information system, where $\mathbb{U} = \{x_1, \ldots, x_i, \ldots, x_n\}$ is the universe of discourse, and $\mathbb{A}$ is a family of attributes, also called knowledge in the universe. V is the value domain of $\mathbb{A}$, and f is an information function $f : \mathbb{U} \times \mathbb{A} \rightarrow V$ [14]. Any subset $\mathbb{P}$ of knowledge $\mathbb{A}$ defines an equivalence or indiscernability relation $IND(\mathbb{P})$ on $\mathbb{U}$, i.e.,

$$IND(\mathbb{P}) = \{(x_i, x_j) \in \mathbb{U} \times \mathbb{U} | \, \forall a \in \mathbb{P}, f_a(x_i) = f_a(x_j)\}.$$

If $(x_i, x_j) \in IND(\mathbb{P})$, then x_i and x_j are indiscernible by attributes from $\mathbb{P}$. The partition of $\mathbb{U}$ generated by $IND(\mathbb{P})$ is denoted as

$$\mathbb{U}/IND(\mathbb{P}) = \{[x_i]_{\mathbb{P}} : x_i \in \mathbb{U}\} \tag{1}$$

where $[x_i]_{\mathbb{P}}$ is the equivalence class containing x_i. The elements in $[x_i]_{\mathbb{P}}$ are indiscernible or equivalent with respect to knowledge $\mathbb{P}$. Equivalence classes, also termed as information granules, are used to characterize arbitrary subsets of $\mathbb{U}$. The equivalence classes of $IND(\mathbb{P})$ and the empty set $\emptyset$ are the elementary sets in the approximation space $\langle \mathbb{U}, \mathbb{A} \rangle$.

Given an arbitrary set $X \subseteq \mathbb{U}$, in general, it may not be possible to describe X precisely in $\langle \mathbb{U}, \mathbb{A} \rangle$. One may characterize X by a pair of lower and upper approximations defined as follows [14]:

$$\underline{\mathbb{P}}(X) = \bigcup \{[x_i]_{\mathbb{P}} \mid [x_i]_{\mathbb{P}} \subseteq X\} \text{ and}$$
$$\overline{\mathbb{P}}(X) = \bigcup \{[x_i]_{\mathbb{P}} \mid [x_i]_{\mathbb{P}} \cap X \neq \emptyset\}.$$

Hence, the lower approximation $\underline{\mathbb{P}}(X)$ is the union of all the elementary sets that are subsets of X, and the upper approximation $\overline{\mathbb{P}}(X)$ is the union of all the elementary sets that have a nonempty intersection with X. The tuple $\langle \underline{\mathbb{P}}(X), \overline{\mathbb{P}}(X) \rangle$ is the representation of an ordinary set X in the approximation space $\langle \mathbb{U}, \mathbb{A} \rangle$ or simply called the rough sets of X. The lower (respectively, upper) approximation $\underline{\mathbb{P}}(X)$ [respectively, $\overline{\mathbb{P}}(X)$] is interpreted as the collection of those elements of $\mathbb{U}$ that definitely (respectively, possibly) belong to X. The lower approximation is also called the positive region sometimes, denoted as $POS_{\mathbb{P}}(X)$. A set X is said to be definable or exact in $\langle \mathbb{U}, \mathbb{A} \rangle$ iff $\underline{\mathbb{P}}(X) = \overline{\mathbb{P}}(X)$. Otherwise, X is indefinable and termed as a rough set.

An information system $\langle \mathbb{U}, \mathbb{A} \rangle$ is called a decision table if the attribute set $\mathbb{A} = \mathbb{C} \cup \mathbb{D}$, where $\mathbb{C}$ and $\mathbb{D}$ represent the condition and decision attribute sets, respectively. The dependence between $\mathbb{C}$ and $\mathbb{D}$ can be defined as

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|POS_{\mathbb{C}}(\mathbb{D})|}{|\mathbb{U}|} \tag{2}$$

where $POS_{\mathbb{C}}(\mathbb{D}) = \cup \underline{\mathbb{C}} X_i$, X_i is the ith equivalence class induced by $\mathbb{D}$, and $|\cdot|$ denotes the cardinality of a set.

A crisp equivalence relation induces a crisp partition of the universe and generates a family of crisp equivalence classes. Correspondingly, a fuzzy equivalence relation generates a fuzzy partition of the universe and a series of fuzzy equivalence classes [15]–[17]. This means that the decision attributes and the condition attributes may all be fuzzy.

Let $\langle \mathbb{U}, \mathbb{A} \rangle$ represent a fuzzy approximation space, and X is a fuzzy subset of $\mathbb{U}$. The fuzzy $\mathbb{P}$-lower and $\mathbb{P}$-upper approximations are then defined as follows [15]:

$$\mu_{\underline{\mathbb{P}}X}(F_i) = \inf_x \left\{ \max \left\{ (1 - \mu_{F_i}(x)), \mu_X(x) \right\} \right\} \qquad \forall i \tag{3}$$

$$\mu_{\overline{\mathbb{P}}X}(F_i) = \sup_x \left\{ \min \left\{ \mu_{F_i}(x), \mu_X(x) \right\} \right\} \qquad \forall i \tag{4}$$

where F_i represents a fuzzy equivalence class belonging to $\mathbb{U}/\mathbb{P}$, and $\mu_X(x)$ represents the membership of x in X. These definitions diverge a little from the crisp upper and lower approximations, as the memberships of individual objects to the approximations are not explicitly available. As a result, the fuzzy lower and upper approximations are defined as [17]

$$\mu_{\underline{\mathbb{P}}X}(x) = \sup_{F_i \in \mathbb{U}/\mathbb{P}} \min \left\{ \mu_{F_i}(x), \mu_{\underline{\mathbb{P}}X}(F_i) \right\} \tag{5}$$

$$\mu_{\overline{\mathbb{P}}X}(x) = \sup_{F_i \in \mathbb{U}/\mathbb{P}} \min \left\{ \mu_{F_i}(x), \mu_{\overline{\mathbb{P}}X}(F_i) \right\}. \tag{6}$$

The tuple $\langle \underline{\mathbb{P}}X, \overline{\mathbb{P}}X \rangle$ is called a fuzzy–rough set. This definition degenerates to traditional rough sets when all equivalence classes are crisp. The membership of an object $x \in \mathbb{U}$ belonging to the fuzzy positive region is

$$\mu_{POS_{\mathbb{C}}(\mathbb{D})}(x) = \sup_{X \in \mathbb{U}/\mathbb{D}} \quad \mu_{\underline{\mathbb{C}}X}(x) \tag{7}$$

where $\mathbb{A} = \mathbb{C} \cup \mathbb{D}$, $\mathbb{C}$ and $\mathbb{D}$ represent the fuzzy condition and decision attribute sets, respectively, and $\mathbb{U}/\mathbb{D}$ represents the partition of $\mathbb{U}$ generated by the decision attribute set $\mathbb{D}$. Using the definition of the fuzzy positive region, the dependence function can be defined as follows [17]:

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{\left|\mu_{POS_{\mathbb{C}}(\mathbb{D})}(x)\right|}{|\mathbb{U}|} = \frac{1}{|\mathbb{U}|} \sum_{x \in \mathbb{U}} \mu_{POS_{\mathbb{C}}(\mathbb{D})}(x). \tag{8}$$

III. Fuzzy-Rough Supervised Attribute Clustering

Here, a new supervised gene clustering algorithm, termed as FRSAC, is presented for grouping coregulated genes with strong association to the class labels. It is based on a supervised similarity measure that follows next.

A. *Fuzzy–Rough Supervised Similarity Measure*

A new quantitative measure, called fuzzy–rough supervised similarity, is defined next based on the definition of the fuzzy positive region of fuzzy–rough sets to compute the similarity between two random variables. It incorporates the information of sample categories or class labels while measuring the similarity between attributes.

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 functional dependence between values of $\mathbb{Q}$ and $\mathbb{P}$, then $\mathbb{Q}$ depends totally on $\mathbb{P}$.

Let $\mathbb{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots, \mathcal{A}_j, \ldots, \mathcal{A}_{\mathcal{D}}\}$ denote the set of $\mathcal{D}$ fuzzy condition attributes of a given data set. Define $\mathrm{R}_{\mathcal{A}_i}(\mathbb{D})$ as the relevance of the fuzzy condition attribute $\mathcal{A}_i$ with respect to the class label or fuzzy decision attribute $\mathbb{D}$. The dependence function of fuzzy–rough sets can be used to calculate the relevance of fuzzy condition attributes. Hence, the relevance $\mathrm{R}_{\mathcal{A}_i}(\mathbb{D})$ of the fuzzy condition attribute $\mathcal{A}_i$ with respect to the fuzzy decision attribute $\mathbb{D}$ using fuzzy–rough sets can be calculated as follows:

$$\mathrm{R}_{\mathcal{A}_i}(\mathbb{D}) = \gamma_{\mathcal{A}_i}(\mathbb{D}) \tag{9}$$

where $\gamma_{\mathcal{A}_i}(\mathbb{D})$ represents the degree of dependence between fuzzy condition attribute $\mathcal{A}_i$ and fuzzy decision attribute or class label $\mathbb{D}$ that is given by (8).

Definition 1: The significance of a fuzzy condition attribute $\mathcal{A}_j$ with respect to another condition attribute $\mathcal{A}_i$ can then be defined as follows:

$$\sigma_{\mathcal{A}_i}(\mathbb{D}, \mathcal{A}_j) = \mathrm{R}_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}) - \mathrm{R}_{\mathcal{A}_i}(\mathbb{D}). \tag{10}$$

Hence, the significance of a fuzzy condition attribute $\mathcal{A}_j$ is the change in dependence when the attribute $\mathcal{A}_j$ is removed from the set $\{\mathcal{A}_i, \mathcal{A}_j\}$. The higher the change in dependence, the more significant the attribute $\mathcal{A}_j$. If the significance is 0, then the attribute $\mathcal{A}_j$ is dispensable.

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

Definition 2: The fuzzy–rough supervised similarity measure between two attributes $\mathcal{A}_i$ and $\mathcal{A}_j$ is defined as follows:

$$\psi(\mathcal{A}_i, \mathcal{A}_j) = 1 - \kappa \tag{11}$$

$$\text{where} \quad \kappa = \left\{ \frac{\sigma_{\mathcal{A}_i}(\mathbb{D}, \mathcal{A}_j) + \sigma_{\mathcal{A}_j}(\mathbb{D}, \mathcal{A}_i)}{2} \right\} \tag{12}$$

$$\text{that is,} \quad \kappa = \mathrm{R}_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}) - \left\{ \frac{\mathrm{R}_{\mathcal{A}_i}(\mathbb{D}) + \mathrm{R}_{\mathcal{A}_j}(\mathbb{D})}{2} \right\}. \tag{13}$$

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, $\psi(\mathcal{A}_i, \mathcal{A}_j) = 0$. 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) $0 \leq \psi(\mathcal{A}_i, \mathcal{A}_j) \leq 1$.
2) $\psi(\mathcal{A}_i, \mathcal{A}_j) = 1$ if and only if $\mathcal{A}_i$ and $\mathcal{A}_j$ are completely correlated.
3) $\psi(\mathcal{A}_i, \mathcal{A}_j) = 0$ 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).

Hence, the fuzzy–rough supervised similarity measure $\psi(\mathcal{A}_i, \mathcal{A}_j)$ between two attributes $\mathcal{A}_i$ and $\mathcal{A}_j$ can be used to compute the redundancy among the attributes taking into account the information of the class label while computing the similarity between two attributes.

Example 1: Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two attributes having relevance values $\gamma_{\mathcal{A}_1}(\mathbb{D}) = 0.375$ and $\gamma_{\mathcal{A}_2}(\mathbb{D}) = 0.381$, respectively, with respect to the class label $\mathbb{D}$. If the joint relevance of these two attributes $\gamma_{\{\mathcal{A}_1, \mathcal{A}_2\}}(\mathbb{D}) = 0.618$, then the significance values of $\mathcal{A}_1$ and $\mathcal{A}_2$ are

$$\sigma_{\mathcal{A}_2}(\mathbb{D}, \mathcal{A}_1) = (0.618 - 0.381) = 0.237$$
$$\sigma_{\mathcal{A}_1}(\mathbb{D}, \mathcal{A}_2) = (0.618 - 0.375) = 0.243$$

respectively, while the supervised similarity between $\mathcal{A}_1$ and $\mathcal{A}_2$ is given by

$$\psi(\mathcal{A}_1, \mathcal{A}_2) = 1 - \left[0.618 - \left(\frac{0.375 + 0.381}{2}\right)\right] = 0.760.$$

B. Proposed Supervised Gene 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 properties of the proposed clustering approach is that the cluster is augmented by attributes that satisfy the 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.

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

Let $\mathrm{R}_{\mathcal{A}_i}(\mathbb{D})$ represent the relevance of attribute $\mathcal{A}_i \in \mathbb{C}$ 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 then formed by selecting the set of attributes $\{\mathcal{A}_j\}$ from the whole set $\mathbb{C}$ considering the attribute $\mathcal{A}_i$ as the representative of cluster $\mathbb{V}_i$, where

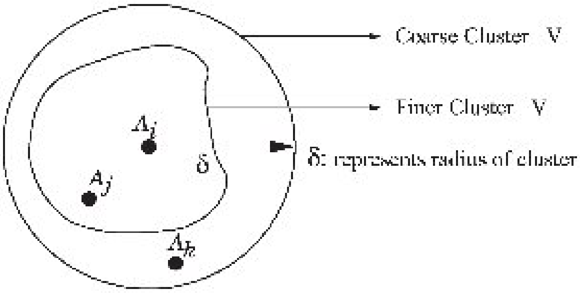


Fig. 1. Representation of a supervised attribute cluster.

$$\mathbb{V}_i = \{\mathcal{A}_j | \psi(\mathcal{A}_i, \mathcal{A}_j) \geq \delta \; \mathcal{A}_j \neq \mathcal{A}_i \in \mathbb{C}\}. \tag{14}$$

Hence, the cluster $\mathbb{V}_i$ represents the set of attributes of $\mathbb{C}$ that 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 $\bar{\mathbb{V}}_i \subset \mathbb{V}_i$ that 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 the coarse cluster reduces the redundancy among attributes of the set $\mathbb{C}$, that of the 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 $\bar{\mathbb{V}}_i$ from the whole set $\mathbb{C}$.

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

Let $\mathbb{C}$ represent the set of attributes of the original data set, while $\mathbb{S}$ and $\bar{\mathbb{S}}$ be the set of actual and augmented attributes, respectively, selected by the proposed attribute clustering algorithm.

Let $\mathbb{V}_i$ be the coarse cluster associated with the attribute $\mathcal{A}_i$, and let $\bar{\mathbb{V}}_i$, which is the finer cluster of $\mathcal{A}_i$ (Fig. 1), represent the set of attributes of $\mathbb{V}_i$ that are merged and averaged with the attribute $\mathcal{A}_i$ to generate the augmented cluster representative $\bar{\mathcal{A}}_i$.

1) Initialize $\mathbb{C} \leftarrow \{\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots, \mathcal{A}_j, \ldots, \mathcal{A}_D\}$, $\mathbb{S} \leftarrow \emptyset$, 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 4–12) until $\mathbb{C} = \emptyset$ or the desired number of attributes is selected.

4) Select attribute $\mathcal{A}_i$ from $\mathbb{C}$ as the representative of cluster $\mathbb{V}_i$ that has the 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 the following four steps (steps 8–11) 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:

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

$$\bar{\mathcal{A}}^{-}_{i+j} = \frac{1}{|\bar{\mathbb{V}}_i| + 1} \left\{ \sum_{\mathcal{A}_k \in \bar{\mathbb{V}}_i} \mathcal{A}_k - \mathcal{A}_j \right\}. \tag{16}$$

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:

$$\mathcal{A}_{i+j} = \begin{cases} \bar{\mathcal{A}}^{+}_{i+j} & \text{if } \mathrm{R}_{\bar{\mathcal{A}}^{+}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathcal{A}}^{-}_{i+j}}(\mathbb{D}) \\ \bar{\mathcal{A}}^{-}_{i+j} & \text{otherwise.} \end{cases} \tag{17}$$

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

In this regard, it can be shown that as the number of desired clusters is constant and sufficiently small compared with the total number of attributes $\mathcal{D}$, the proposed clustering algorithm has an overall $\mathcal{O}(\mathcal{D})$ time complexity.

C. Fundamental Property

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

1) $\psi(\mathcal{A}_i, \mathcal{A}_j) \geq \delta; \forall \mathcal{A}_j \in \mathbb{V}_i$;
2) $\mathrm{R}_{\mathcal{A}_i}(\mathbb{D}) \geq \mathrm{R}_{\mathcal{A}_j}(\mathbb{D}); \forall \mathcal{A}_j \in \mathbb{V}_i$;
3) $\mathrm{R}_{\bar{\mathcal{A}}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathcal{A}}_i}(\mathbb{D}); \forall \mathcal{A}_j \in \bar{\mathbb{V}}_i$;
4) $\mathrm{R}_{\bar{\mathcal{A}}_{i+j}}(\mathbb{D}) < \mathrm{R}_{\bar{\mathcal{A}}_i}(\mathbb{D}); \forall \mathcal{A}_j \in \mathbb{V}_i \setminus \bar{\mathbb{V}}_i$.

Property 1 says that if an attribute $\mathcal{A}_j \in \mathbb{V}_i \Rightarrow \psi(\mathcal{A}_i, \mathcal{A}_j) \geq \delta$. That is, the supervised similarity between the attribute $\mathcal{A}_j$ of coarse cluster $\mathbb{V}_i$ and the initial cluster representative $\mathcal{A}_i$ is greater than a predefined threshold value δ. Property 2 establishes the fact that if $\mathcal{A}_j \in \mathbb{V}_i \Rightarrow \mathrm{R}_{\mathcal{A}_i}(\mathbb{D}) \geq \mathrm{R}_{\mathcal{A}_j}(\mathbb{D})$, that is, the relevance of the cluster representative $\mathcal{A}_i$ is the maximum among that of all attributes of the cluster $\mathbb{V}_i$. Properties 3 and 4 are of great importance in increasing the relevance of the augmented cluster representative with respect to the class labels and reducing the redundancy among the attribute set. Property 3 says that if $\mathcal{A}_j \in \bar{\mathbb{V}}_i \Rightarrow \mathrm{R}_{\bar{\mathcal{A}}_{i+j}}(\mathbb{D}) \geq \mathrm{R}_{\bar{\mathcal{A}}_i}(\mathbb{D})$. It means that an attribute $\mathcal{A}_j$ belongs to the finer cluster $\bar{\mathbb{V}}_i$ 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 that belong to only coarse cluster $\mathbb{V}_i$, not to finer cluster $\bar{\mathbb{V}}_i$, are not responsible in increasing the relevance of the augmented cluster representative. Hence, the set of attributes $\bar{\mathbb{V}}_i$ increases the relevance value of the attribute $\mathcal{A}_i$ and reduces the redundancy of the whole set, while the set of attributes $\mathbb{V}_i \setminus \bar{\mathbb{V}}_i$ is only responsible for reducing the redundancy.

D. Generation of Fuzzy Equivalence Classes

The family of normal fuzzy sets produced by fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes. Given a finite set $\mathbb{U}$, $\mathbb{C}$ is a fuzzy condition attribute set in $\mathbb{U}$, which generates a fuzzy equivalence partition on $\mathbb{U}$. If c denotes the number of fuzzy equivalence classes generated by the fuzzy equivalence relation and n is the number of objects in $\mathbb{U}$, then c-partitions of $\mathbb{U}$ are sets of (cn) values $\{\mu^{\mathbb{C}}_{ij}\}$ that can be conveniently arrayed as a $(c \times n)$ matrix $\mathbb{M}_{\mathbb{C}} = [\mu^{\mathbb{C}}_{ij}]$, which is denoted by

$$\mathbb{M}_{\mathbb{C}} = \begin{pmatrix} \mu^{\mathbb{C}}_{11} & \mu^{\mathbb{C}}_{12} & \cdots & \mu^{\mathbb{C}}_{1n} \\ \mu^{\mathbb{C}}_{21} & \mu^{\mathbb{C}}_{22} & \cdots & \mu^{\mathbb{C}}_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \mu^{\mathbb{C}}_{c1} & \mu^{\mathbb{C}}_{c2} & \cdots & \mu^{\mathbb{C}}_{cn} \end{pmatrix} \tag{18}$$

subject to $\sum_{i=1}^{c} \mu^{\mathbb{C}}_{ij} = 1, \forall j$, and for any value of i, if

$$k = \arg\max_{j} \left\{\mu^{\mathbb{C}}_{ij}\right\}, \text{ then } \max_{j} \left\{\mu^{\mathbb{C}}_{ij}\right\} = \max_{l} \left\{\mu^{\mathbb{C}}_{lk}\right\} > 0$$

where $\mu^{\mathbb{C}}_{ij} = \mu_{F_i}(x_j) \in [0, 1]$ represents the membership of object x_j in the ith fuzzy equivalence class F_i. In the proposed supervised gene-clustering method, the π function in the 1-D form is used to assign membership values to different fuzzy equivalence classes for the input genes. A fuzzy set with membership function $\pi(x; \bar{c}, \sigma)$ represents a set of points clustered around $\bar{c}$, where

$$\pi(x\bar{c}, \sigma) = \begin{cases} 2\left(1 - \frac{\|x-\bar{c}\|}{\sigma}\right)^2 & \text{for } \frac{\sigma}{2} \leq \|x - \bar{c}\| \leq \sigma \\ 1 - 2\left(\frac{\|x-\bar{c}\|}{\sigma}\right)^2 & \text{for } 0 \leq \|x - \bar{c}\| \leq \frac{\sigma}{2} \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

where $\sigma > 0$ is the radius of the π function, with $\bar{c}$ as the central point, and $\| \cdot \|$ denotes the Euclidean norm [22]. When the pattern x lies at the central point $\bar{c}$ of a class, then $\|x - \bar{c}\| = 0$, and its membership value is maximum, that is, $\pi(\bar{c}; \bar{c}, \sigma) = 1$. The membership value of a point decreases as its distance from the central point $\bar{c}$, that is, $\|x - \bar{c}\|$, increases. When $\|x - \bar{c}\| = (\sigma/2)$, the membership value of x is 0.5, and this is called a crossover point. The $(c \times n)$ matrix $\mathbb{M}_{\mathcal{A}_i}$, corresponding to the

ith gene $\mathcal{A}_i$, can be calculated from the c-fuzzy equivalence classes of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$, where

$$\mu^{\mathcal{A}_i}_{kj} = \frac{\pi(x_j \bar{c}_k, \sigma_k)}{\sum\limits_{l=1}^{c} \pi(x_j \bar{c}_l, \sigma_l)}. \tag{20}$$

Each input real-valued gene in a quantitative form can be assigned to different fuzzy equivalence classes in terms of membership values using the π fuzzy set with appropriate $\bar{c}$ and σ. The centers and the radii of the π functions along each gene axis are determined from the distribution of the training patterns. In the proposed gene-clustering algorithm, three fuzzy equivalence classes ($c = 3$), namely, low, medium, and high, are considered. These three equivalence classes correspond to underexpression, baseline, and overexpression of continuous-valued genes, respectively. Corresponding to three fuzzy sets low, medium, and high, the following relations hold:

$$\begin{aligned} \bar{c}_1 &= \bar{c}_{\text{low}}(\mathcal{A}_i)\ \bar{c}_2 = \bar{c}_{\text{medium}}(\mathcal{A}_i)\ \bar{c}_3 = \bar{c}_{\text{high}}(\mathcal{A}_i) \\ \sigma_1 &= \sigma_{\text{low}}(\mathcal{A}_i)\ \sigma_2 = \sigma_{\text{medium}}(\mathcal{A}_i)\ \sigma_3 = \sigma_{\text{high}}(\mathcal{A}_i). \end{aligned}$$

The parameters $\bar{c}$ and σ of each π fuzzy set are computed according to the following procedure [22]. Let $\bar{m}_i$ be the mean of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$ along the ith gene $\mathcal{A}_i$. Then, $\bar{m}_{i_l}$ and $\bar{m}_{i_h}$ are defined as the mean along the ith gene of the objects having coordinate values in the range $[\mathcal{A}_{i_{\min}}, \bar{m}_i)$ and $(\bar{m}_i, \mathcal{A}_{i_{\max}}]$, respectively, where $\mathcal{A}_{i_{\max}}$ and $\mathcal{A}_{i_{\min}}$ denote the upper and lower bounds of the dynamic range of gene $\mathcal{A}_i$ for the training set. For three fuzzy sets low, medium, and high, the centers and the corresponding radii are computed as follows:

$$\begin{aligned} \bar{c}_{\text{low}}(\mathcal{A}_i) &= \bar{m}_{i_l}\ \bar{c}_{\text{medium}}(\mathcal{A}_i) = \bar{m}_i\ \bar{c}_{\text{high}}(\mathcal{A}_i) = \bar{m}_{i_h} \\ \sigma_{\text{low}}(\mathcal{A}_i) &= 2\left(\bar{c}_{\text{medium}}(\mathcal{A}_i) - \bar{c}_{\text{low}}(\mathcal{A}_i)\right) \\ \sigma_{\text{high}}(\mathcal{A}_i) &= 2\left(\bar{c}_{\text{high}}(\mathcal{A}_i) - \bar{c}_{\text{medium}}(\mathcal{A}_i)\right) \\ \sigma_{\text{medium}}(\mathcal{A}_i) &= \eta \times \frac{\text{A}}{\text{B}} \end{aligned}$$

where

$$\begin{aligned} \text{A} = \{&\sigma_{\text{low}}(\mathcal{A}_i)\,(\mathcal{A}_{i_{\max}} - c_{\text{medium}}(\mathcal{A}_i)) \\ &+ \sigma_{\text{high}}(\mathcal{A}_i)\,(c_{\text{medium}}(\mathcal{A}_i) - \mathcal{A}_{i_{\min}})\} \\ \text{B} = \{&\mathcal{A}_{i_{\max}} - \mathcal{A}_{i_{\min}}\} \end{aligned}$$

where η is a multiplicative parameter controlling the overlapping. The distribution of the patterns along each gene axis is taken into account, while computing the corresponding centers and radii of the fuzzy sets. Also, the amount of overlap between the three fuzzy sets can be different along the different axis, depending on the distribution of the patterns.

IV. Experimental Results

The performance of the proposed FRSAC algorithm is compared with that of some existing supervised and unsupervised gene-clustering and gene-selection algorithms, namely, the fuzzy equivalence partition matrix (FEPM)-based gene selection algorithm [21], rough–fuzzy c-means (RFCM) [18], the supervised gene clustering algorithm (SGCA) [10], the attribute clustering algorithm (ACA) [5], fuzzy c-means (FCM) [20], and the minimum redundancy–maximum relevance (mRMR) framework [23]. To analyze the performance of different algorithms, the experimentation is done on eight microarray gene expression data sets, namely, breast, leukemia, colon, prostate, lung, RBreast, RAOA, and RAHC data sets [21]; each of the data is preprocessed by standardizing each sample to zero mean and unit variance. The source code of the FRSAC algorithm written in C language and the supplementary information are available at http://www.isical.ac.in/~pmaji/results/frsac.html.

A. Class Prediction Methods

The major metrics for evaluating the performance are the CS index [2] and the classification accuracy of the NB classifier [2], the K-NN rule [2], and the SVM [24]. To compute the classification accuracy of three classifiers, the leave-one-out cross-validation is performed on each data set.

1) SVM: The SVM [24] 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 a nonlinear decision boundary. In this paper, 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 [2] is used for evaluating the effectiveness of the reduced feature set for classification. It classifies samples based on the 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-NNs. The value of K, chosen for the K-NN, is the square root of the number of samples in the training set.

3) NB Classifier: The NB classifier [2] is obtained 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 $\{w_{1j}, \ldots, w_{ij}, \ldots, w_{mj}\}$ for m genes, the posterior probability that x_j belongs to class c is

$$p(c|x_j) \propto \prod_{i=1}^{m} p(w_{ij}|c) \tag{21}$$

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

4) CS Index: The CS index $\mathcal{S}$ [2] of a data set is defined as $\mathcal{S} = \text{trace}(S_b^{-1} S_w)$, where S_w and S_b are the within and between class scatter matrices, respectively, defined as follows:

$$\begin{aligned} S_w &= \sum_{j=1}^{C} \pi_j E\left\{(X - \mu_j)(X - \mu_j)^T | c_j\right\} = \sum_{j=1}^{C} \pi_j \Sigma_j \\ S_b &= \sum_{j=1}^{C} (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^T \text{ and } \bar{\mu} = E\{X\} = \sum_{j=1}^{C} \pi_j \mu_j \end{aligned}$$

where C is the number of classes, π_j is the *a priori* probability that a pattern belongs to class c_j, X is a feature vector,

TABLE I
PERFORMANCE OF THE PROPOSED ALGORITHM ON COLON AND LUNG CANCER DATA SETS FOR $0.90 \le \delta \le 0.96$ AND $1.0 \le \eta \le 1.5$

Value of η	Value of m	Measure	Colon Cancer / Different Values of Threshold δ							Lung Cancer / Different Values of Threshold δ						
			0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.90	0.91	0.92	0.93	0.94	0.95	0.96
1.0	1	SVM	95.2	93.5	91.9	88.7	88.7	79.0	80.6	100	100	98.3	99.4	99.4	98.9	98.9
		K-NN	93.6	98.4	90.3	82.3	82.3	79.0	79.0	99.4	100	98.3	99.4	99.4	98.9	98.9
	2	SVM	98.4	100	95.2	87.1	85.5	83.9	82.3	100	100	100	100	100	100	100
		K-NN	96.8	98.4	95.2	87.1	93.6	79.0	83.9	100	100	99.4	100	100	98.3	98.9
	3	SVM	100	100	95.2	91.9	87.1	80.6	85.5	100	100	100	100	100	100	100
		K-NN	98.4	100	91.9	91.9	87.1	85.5	83.9	100	100	100	99.4	99.4	98.9	98.9
1.1	1	SVM	95.2	100	96.8	93.5	74.2	80.6	83.9	99.4	99.4	99.4	99.4	99.4	99.4	100
		K-NN	100	100	100	91.9	75.8	75.8	79.0	98.9	98.9	98.9	98.9	98.9	98.9	100
	2	SVM	100	100	100	93.5	75.8	82.3	83.9	98.9	98.9	98.9	98.9	98.9	98.9	99.4
		K-NN	100	100	100	95.2	85.5	82.3	80.7	98.9	98.9	98.9	98.9	98.9	98.9	99.4
	3	SVM	100	100	100	96.8	75.8	82.3	79.0	99.4	99.4	99.4	99.4	99.4	99.4	99.4
		K-NN	100	100	98.4	95.2	71.0	82.3	79.0	98.9	98.9	98.9	98.9	98.9	98.9	99.4
1.2	1	SVM	96.8	95.2	98.4	80.6	77.4	77.4	82.3	99.4	99.4	99.4	100	100	100	100
		K-NN	96.8	95.2	98.4	82.3	74.2	75.8	72.6	99.4	99.4	99.4	100	100	99.4	100
	2	SVM	95.2	95.2	95.2	91.9	75.8	79.0	82.3	100	100	100	100	100	100	100
		K-NN	96.8	95.2	95.2	88.7	88.7	71.0	79.0	100	100	100	100	100	100	100
	3	SVM	95.2	98.4	91.9	95.2	87.1	79.0	87.1	100	100	100	100	100	100	100
		K-NN	96.8	98.4	95.2	95.2	87.1	66.1	83.9	100	100	100	100	100	100	99.4
1.3	1	SVM	96.8	96.8	96.8	96.8	93.5	96.8	79.0	82.9	82.9	82.9	82.9	82.9	82.9	82.9
		K-NN	98.4	98.4	98.4	98.4	98.4	96.8	83.9	91.2	91.2	91.2	91.2	91.2	91.2	91.2
	2	SVM	100	95.2	96.8	91.9	95.2	96.8	90.3	99.4	99.4	99.4	100	100	82.9	82.9
		K-NN	100	96.8	96.8	95.2	95.2	91.9	91.9	99.4	99.4	99.4	100	100	95.0	95.0
	3	SVM	98.4	96.8	95.2	93.5	95.2	95.2	82.3	100	100	100	100	100	99.4	99.4
		K-NN	100	96.8	98.4	96.8	93.6	93.6	82.3	100	100	100	100	100	100	100
1.4	1	SVM	95.2	95.2	95.2	93.5	98.4	93.5	87.1	82.9	82.9	82.9	82.9	82.9	82.9	82.9
		K-NN	95.2	95.2	95.2	95.2	98.4	98.4	87.1	93.4	93.4	93.4	93.4	93.4	93.4	93.4
	2	SVM	93.5	93.5	93.5	91.9	96.8	95.2	87.1	82.9	82.9	82.9	82.9	82.9	82.9	82.9
		K-NN	95.2	95.2	95.2	95.2	93.6	93.6	90.3	93.4	93.4	93.4	93.4	93.4	93.4	93.4
	3	SVM	90.3	90.3	90.3	96.8	96.8	93.5	90.3	100	100	100	100	100	100	82.9
		K-NN	91.9	91.9	91.9	95.2	95.2	93.6	88.7	100	100	100	100	100	100	93.9
1.5	1	SVM	88.7	88.7	88.7	88.7	96.8	87.1	93.5	100	100	100	100	100	100	100
		K-NN	87.1	87.1	87.1	87.1	100	90.3	91.9	100	100	100	100	100	100	100
	2	SVM	88.7	88.7	88.7	88.7	100	90.3	90.3	100	100	100	100	100	100	100
		K-NN	91.9	91.9	91.9	91.9	98.4	91.9	90.3	100	100	100	100	100	100	100
	3	SVM	90.3	90.3	90.3	90.3	98.4	91.9	90.3	100	100	100	100	100	100	100
		K-NN	91.9	91.9	91.9	91.9	96.8	91.9	85.5	100	100	100	100	100	100	100

$\bar{\mu}$ is the sample mean vector for the entire data points, μ_j and Σ_j represent the sample mean and the covariance matrix of class c_j, and $E\{\cdot\}$ is the expectation operator. A lower value of S ensures that classes are well separated by their scatter means.

B. Role of the Threshold and Multiplicative Parameter

The threshold δ in (14) 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. The similarity among the attributes of a cluster decreases with the decrease in the value of δ. On the other hand, the multiplicative parameter η controls the overlapping between three fuzzy equivalence classes low and medium or medium and high. Keeping the values of σ_{low} and σ_{high} fixed, the amount of overlapping among the three π functions can be altered varying σ_{medium}. As η is decreased, the radius σ_{medium} decreases around $\bar{c}_{\mathrm{medium}}$ such that, ultimately, there is insignificant overlapping between the π functions low and medium or medium and high. As η increases, the radius σ_{medium} increases around $\bar{c}_{\mathrm{medium}}$ such that the amount of overlapping between π functions increases.

To find out the optimum values of both η and δ, the extensive experimentation is carried out on eight microarray data sets. Tables I and II represent the performance of the proposed clustering algorithm on colon, lung, prostate cancer, and RBreast data sets for different values of η. The results and subsequent discussions are presented with respect to the classification accuracy of the SVM and the K-NN rule, and $0.90 \le \delta \le 0.96$. The results are reported for three best clusters $(m = 3)$ obtained using the proposed attribute-clustering method. From the results reported in Tables I and II, it is seen that very large or very small amounts of overlapping among the three fuzzy equivalence classes of the input feature are found to be undesirable irrespective of the values of δ. The proposed supervised attribute clustering algorithm achieves its best performance at $\eta = 1.1$ for colon cancer data, 1.2 for lung and prostate cancer data, 1.3 for RBreast data, and 1.0 and 1.5 for lung cancer data. That is, the best performance of the proposed algorithm is obtained on these four data sets for $1.1 \le \eta \le 1.3$ with respect to the classification accuracy of the SVM and the K-NN rule.

Table III represents the performance of the proposed supervised attribute clustering algorithm on colon, lung, prostate cancer, and RBreast data sets for different values of δ. The results and subsequent discussions are presented with respect to the classification accuracy of the SVM and the K-NN rule, and $1.1 \le \eta \le 1.3$. From the results reported in Table III, it is seen that as the value of δ increases, the classification accuracy of the

TABLE II

Performance of the Proposed Algorithm on RBreast and Prostate Cancer Data Sets for $0.90 \le \delta \le 0.96$ and $1.0 \le \eta \le 1.5$

Value of η	Value of m	Measure	RBreast / Different Values of Threshold δ							Prostate Cancer / Different Values of Threshold δ						
			0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.90	0.91	0.92	0.93	0.94	0.95	0.96
1.0	1	SVM	63.9	63.9	63.9	63.9	63.9	63.9	63.9	90.4	86.8	87.5	91.2	87.5	78.7	64.0
		K-NN	75.3	75.3	75.3	75.3	75.3	75.3	75.3	96.3	85.3	96.3	95.6	83.1	87.5	75.0
	2	SVM	62.9	62.9	62.9	62.9	62.9	62.9	62.9	93.4	90.4	97.1	92.6	88.2	88.2	64.0
		K-NN	62.9	62.9	62.9	62.9	62.9	62.9	62.9	92.7	93.4	97.8	97.1	91.2	90.4	83.8
	3	SVM	61.9	61.9	61.9	61.9	61.9	61.9	61.9	94.1	91.9	96.3	94.1	89.0	90.4	66.2
		K-NN	72.2	72.2	72.2	72.2	72.2	72.2	72.2	94.1	91.9	98.5	97.8	94.8	94.8	81.6
1.1	1	SVM	64.9	64.9	64.9	64.9	64.9	64.9	64.9	94.9	94.9	96.3	85.3	89.7	81.6	81.6
		K-NN	83.5	83.5	83.5	83.5	83.5	83.5	83.5	100	100	100	93.4	89.7	85.3	86.8
	2	SVM	64.9	64.9	64.9	64.9	64.9	64.9	64.9	100	100	100	91.9	94.1	91.2	83.1
		K-NN	73.2	73.2	73.2	73.2	73.2	73.2	73.2	100	100	100	96.3	93.4	87.5	87.5
	3	SVM	64.9	64.9	64.9	64.9	64.9	64.9	64.9	100	100	100	94.9	93.4	91.9	83.1
		K-NN	69.1	69.1	69.1	69.1	69.1	69.1	69.1	100	100	100	94.1	93.4	86.0	87.5
1.2	1	SVM	64.9	64.9	64.9	64.9	64.9	64.9	64.9	86.8	91.2	91.2	100	91.2	89.7	80.1
		K-NN	75.3	75.3	75.3	75.3	75.3	75.3	75.3	98.5	96.3	97.1	100	94.1	88.2	88.2
	2	SVM	73.2	73.2	73.2	73.2	73.2	73.2	64.9	94.1	94.1	94.9	100	94.1	92.6	90.4
		K-NN	72.2	72.2	72.2	72.2	72.2	72.2	74.2	99.3	97.1	99.3	100	94.1	95.6	93.4
	3	SVM	73.2	73.2	73.2	73.2	73.2	73.2	67.0	97.8	97.1	95.6	100	95.6	91.2	89.7
		K-NN	74.2	74.2	74.2	74.2	74.2	74.2	72.2	96.3	97.1	96.3	100	94.8	92.7	91.2
1.3	1	SVM	77.3	77.3	77.3	77.3	77.3	77.3	99.0	88.2	78.7	83.1	83.1	86.0	82.4	75.7
		K-NN	83.5	83.5	83.5	83.5	83.5	83.5	100	89.7	82.3	89.7	76.5	84.6	86.0	75.7
	2	SVM	89.7	89.7	89.7	89.7	89.7	89.7	99.0	91.2	87.5	89.7	89.7	86.0	89.0	80.1
		K-NN	87.6	87.6	87.6	87.6	87.6	87.6	99.0	91.9	89.7	89.7	88.2	90.4	85.3	85.3
	3	SVM	90.7	90.7	90.7	90.7	90.7	90.7	100	92.6	90.4	89.7	90.4	85.3	90.4	83.8
		K-NN	86.6	86.6	86.6	86.6	86.6	86.6	100	91.9	89.0	87.5	86.8	85.3	85.3	88.2
1.4	1	SVM	89.7	89.7	89.7	89.7	77.3	80.4	78.4	86.8	81.6	89.7	88.2	80.9	85.3	86.0
		K-NN	99.0	99.0	99.0	99.0	83.5	78.3	79.4	92.7	90.4	81.6	83.8	81.6	86.0	80.2
	2	SVM	99.0	99.0	99.0	99.0	91.8	89.7	87.6	91.9	88.2	87.5	91.9	90.4	89.0	90.4
		K-NN	96.9	96.9	96.9	96.9	91.8	87.6	84.5	96.3	91.9	93.4	94.1	91.2	89.0	89.0
	3	SVM	96.9	96.9	96.9	96.9	86.6	89.7	90.7	91.2	88.2	89.7	90.4	91.9	89.7	91.2
		K-NN	95.9	95.9	95.9	95.9	87.6	92.8	86.6	94.8	93.4	93.4	94.8	95.6	89.0	94.1
1.5	1	SVM	90.7	90.7	90.7	90.7	90.7	90.7	92.8	80.9	80.9	80.9	80.9	79.4	79.4	77.9
		K-NN	92.8	92.8	92.8	89.7	89.7	89.7	91.8	91.2	88.2	77.2	80.9	83.1	83.1	82.3
	2	SVM	99.0	99.0	99.0	99.0	99.0	99.0	94.8	82.4	83.8	84.6	84.6	81.6	81.6	80.9
		K-NN	95.9	95.9	95.9	94.8	94.8	94.8	95.9	89.0	84.6	84.6	82.3	83.1	83.1	82.3
	3	SVM	100	100	100	100	100	100	100	82.4	83.1	84.6	83.8	81.6	83.8	82.4
		K-NN	97.9	97.9	97.9	97.9	97.9	97.9	96.9	85.3	86.8	88.2	85.3	83.1	89.0	87.5

TABLE III

Performance of the Proposed Algorithm for Different Values of Threshold δ

Data Sets	Value of m	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
Colon $\eta = 1.1$	1	SVM	95.2	95.2	95.2	95.2	95.2	95.2	100	96.8	93.5	74.2	80.6	83.9	59.7	61.3	64.5
		K-NN	100	100	100	100	100	100	100	100	91.9	75.8	75.8	79.0	69.3	69.3	72.6
	2	SVM	100	100	100	100	100	100	100	100	93.5	75.8	82.3	83.9	75.8	58.1	67.7
		K-NN	100	100	100	100	100	100	100	100	95.2	85.5	82.3	80.7	61.3	79.0	66.1
	3	SVM	100	100	100	100	100	100	100	100	96.8	75.8	82.3	79.0	79.0	71.0	75.8
		K-NN	100	100	100	100	100	100	100	98.4	95.2	71.0	82.3	79.0	82.3	75.8	71.0
Lung $\eta = 1.2$	1	SVM	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	100	100	100	100	99.4	98.9	98.9
		K-NN	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	100	100	99.4	100	99.4	98.9	98.9
	2	SVM	100	100	100	100	100	100	100	100	100	100	100	100	99.4	100	98.9
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	98.9	98.9
	3	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	99.4	99.4	98.3	98.9
Prostate $\eta = 1.2$	1	SVM	94.9	94.9	90.4	90.4	94.9	86.8	91.2	91.2	100	91.2	89.7	80.1	61.0	61.8	61.0
		K-NN	97.8	97.8	98.5	98.5	97.1	98.5	96.3	97.1	100	94.1	88.2	88.2	73.5	72.8	77.2
	2	SVM	96.3	96.3	95.6	95.6	95.6	94.1	94.1	94.9	100	94.1	92.6	90.4	68.4	66.2	61.8
		K-NN	100	100	100	99.3	98.5	99.3	97.1	99.3	100	94.1	95.6	93.4	81.6	79.4	78.7
	3	SVM	98.5	98.5	98.5	98.5	98.5	97.8	97.1	95.6	100	95.6	91.2	89.7	66.2	69.1	62.5
		K-NN	97.8	97.8	97.1	97.8	98.5	96.3	97.1	96.3	100	94.8	92.7	91.2	82.3	81.6	83.8
RBreast $\eta = 1.3$	1	SVM	77.3	77.3	77.3	77.3	77.3	77.3	77.3	77.3	77.3	77.3	77.3	99.0	99.0	76.3	70.1
		K-NN	83.5	83.5	83.5	83.5	83.5	83.5	83.5	83.5	83.5	83.5	83.5	100	100	85.6	69.1
	2	SVM	89.7	89.7	89.7	89.7	89.7	89.7	89.7	89.7	89.7	89.7	89.7	99.0	96.9	83.5	79.4
		K-NN	87.6	87.6	87.6	87.6	87.6	87.6	87.6	87.6	87.6	87.6	87.6	99.0	96.9	87.6	76.3
	3	SVM	90.7	90.7	90.7	90.7	90.7	90.7	90.7	90.7	90.7	90.7	90.7	100	96.9	81.4	82.5
		K-NN	86.6	86.6	86.6	86.6	86.6	86.6	86.6	86.6	86.6	86.6	86.6	100	96.9	84.5	79.4

SVM and the K-NN rule increases. The proposed supervised attribute clustering algorithm achieves its best performance at $\delta = 0.91$ for colon cancer data, 0.93 for lung and prostate cancer data, 0.94 for lung cancer data, and 0.96 for RBreast data. That is, the algorithm performs best for $0.90 \le \delta \le 0.96$ with respect to the classification accuracy of the SVM and the

TABLE IV
PERFORMANCE OF THE PROPOSED ALGORITHM ON BREAST CANCER AND LEUKEMIA DATA SETS FOR $0.90 \leq \delta \leq 0.96$ AND $1.1 \leq \eta \leq 1.3$

Value of η	Value of m	Measure	Breast Cancer / Different Values of Threshold δ							Leukemia / Different Values of Threshold δ						
			0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.90	0.91	0.92	0.93	0.94	0.95	0.96
1.1	1	SVM	100	100	100	100	87.8	93.9	93.9	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	98.6	100	100	100	100	100
	2	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	98.6
	3	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	98.0	98.0	98.0	100	100	98.0	100	100	100	100	100	100	100	98.6
1.2	1	SVM	100	100	100	100	100	100	100	100	100	100	98.6	100	100	98.6
		K-NN	100	100	100	100	100	100	100	100	100	100	98.6	100	100	98.6
	2	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	98.6
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	98.6
	3	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	98.0	100	100	98.0	98.0	100	100	100	100	100	100	100
1.3	1	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	2	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	3	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	100

TABLE V
PERFORMANCE OF THE PROPOSED ALGORITHM ON RAOA AND RAHC DATA SETS FOR $0.90 \leq \delta \leq 0.96$ AND $1.1 \leq \eta \leq 1.3$

Value of η	Value of m	Measure	RAOA / Different Values of Threshold δ							RAHC / Different Values of Threshold δ						
			0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.90	0.91	0.92	0.93	0.94	0.95	0.96
1.1	1	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	2	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	3	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	100
1.2	1	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	2	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	3	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	100
1.3	1	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	2	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	3	SVM	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		K-NN	100	100	100	100	100	100	100	100	100	100	100	100	100	100

K-NN rule. Finally, Tables IV and V report the performance of the proposed algorithm on breast cancer, leukemia, RAOA, and RAHC data sets for $1.1 \leq \eta \leq 1.3$ and $0.90 \leq \delta \leq 0.96$. Out of 504 cases, the proposed supervised attribute clustering algorithm attains 100% classification accuracy in 485 cases, that is, 96.2% cases on these four data sets. All the results reported in Tables I–V establish the fact that the best performance of the proposed algorithm is achieved for $1.1 \leq \eta \leq 1.3$ and $0.90 \leq \delta \leq 0.96$ irrespective of the data sets used. In other words, the proposed algorithm attains its best performance when the genes are grouped based on at least 90% of their fuzzy–rough supervised similarity values and when at least one of the membership values in any one of the three fuzzy equivalence classes is greater than 0.5.

C. *Importance of Augmented Genes*

Each coarse cluster represents the set of genes that have the fuzzy–rough 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 attribute clustering algorithm. By searching among the genes of the coarse cluster, the current cluster representative is merged and averaged with one single gene 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 the augmented cluster representative of the finer cluster over the initial cluster representative, that is, the actual gene, the extensive experiments are carried out on eight microarray data sets. Table VI reports the comparative performance of actual and augmented genes of different finer clusters as well as that of augmented genes of coarse and finer clusters. Results are reported for $m = 3$ considering fuzzy–rough supervised similarity measure. The performance is compared with respect to the CS index and the classification accuracy of the SVM, the K-NN rule, and the NB classifier. All the results reported in Table VI establish that the proposed supervised attribute clustering algorithm performs significantly better in case of the augmented gene computed from the finer cluster than that of the coarse cluster and the actual gene irrespective of the data sets and quantitative indexes used.

TABLE VI
COMPARATIVE PERFORMANCE ANALYSIS OF AUGMENTED AND ACTUAL GENES FOR DIFFERENT MICROARRAY DATA SETS

Microarray Data Sets	Genes / Attributes	$m = 1$				$m = 2$				$m = 3$			
		SVM	K-NN	NB	CS	SVM	K-NN	NB	CS	SVM	K-NN	NB	CS
Breast	Finer	100	100	100	0.46	100	100	100	0.47	100	100	100	0.45
$\eta = 1.2$	Coarse	51.0	30.6	32.7	68.54	51.0	38.8	65.3	22.00	61.2	44.9	55.1	33.39
$\delta = 0.93$	Actual	69.4	46.9	69.4	4.46	81.6	75.5	77.6	4.91	79.6	81.6	69.4	4.37
Leukemia	Finer	100	100	100	0.47	100	100	100	0.47	100	100	100	0.46
$\eta = 1.2$	Coarse	61.1	68.1	75.0	7.24	75.0	69.4	75.0	15.33	72.2	61.1	75.0	21.24
$\delta = 0.92$	Actual	93.1	91.7	76.4	0.87	95.8	94.4	93.1	0.91	94.4	93.1	88.9	0.94
Colon	Finer	100	100	95.2	0.62	100	100	100	0.59	100	100	98.4	0.63
$\eta = 1.1$	Coarse	64.5	67.7	16.1	403.59	64.5	58.1	61.3	306.90	64.5	54.8	58.1	118.22
$\delta = 0.91$	Actual	72.6	61.3	75.8	4.76	77.4	72.6	80.7	4.11	82.3	79.0	77.4	4.00
Lung	Finer	100	100	98.9	1.28	100	100	100	1.02	100	100	100	0.98
$\eta = 1.2$	Coarse	82.9	82.9	53.6	72.16	82.9	85.1	54.1	16.47	86.2	85.1	77.9	11.29
$\delta = 0.93$	Actual	82.9	82.9	33.7	22.17	82.9	85.1	65.2	17.71	82.9	88.4	80.7	17.51
RBreast	Finer	99.0	100	60.8	0.58	99.0	99.0	89.7	0.58	100	100	95.9	0.71
$\eta = 1.3$	Coarse	52.6	50.5	53.6	83.47	53.6	40.2	53.6	83.47	53.6	48.5	53.6	83.41
$\delta = 0.96$	Actual	56.7	49.5	54.6	160.33	56.7	50.5	55.7	71.31	56.7	42.3	53.6	107.46
Prostate	Finer	100	100	75.0	1.44	100	100	78.7	1.51	100	100	80.2	1.59
$\eta = 1.2$	Coarse	55.1	50.7	59.6	25.09	52.9	57.4	47.8	28.58	53.7	55.9	56.6	30.71
$\delta = 0.93$	Actual	71.3	80.9	84.6	4.19	74.3	80.9	83.8	5.10	74.3	81.6	83.8	6.88
RAOA	Finer	100	100	100	0.52	100	100	100	0.58	100	100	100	0.58
$\eta = 1.2$	Coarse	86.7	80.0	86.7	0.89	96.7	93.3	90.0	0.88	93.3	96.7	90.0	1.73
$\delta = 0.93$	Actual	93.3	90.0	93.3	0.87	90.0	90.0	86.7	1.41	96.7	96.7	93.3	1.20
RAHC	Finer	100	100	100	0.61	100	100	100	0.57	100	100	100	0.62
$\eta = 1.2$	Coarse	70.0	66.0	30.0	4576.50	70.0	64.0	66.0	117.94	68.0	66.0	64.0	31.40
$\delta = 0.93$	Actual	70.0	62.0	44.0	7.07	70.0	70.0	54.0	7.59	70.0	66.0	54.0	7.89

TABLE VII
p-VALUE AND FDR (IN PERCENT) FOR GENES IN BEST CLUSTERS OBTAINED BY DIFFERENT METHODS

Gene Ontology	Methods / Algorithms	Breast		Leukemia		RBreast		Prostate		RAOA	
		p-value	FDR	p-value	FDR	p-value	FDR	p-value	FDR	p-value	FDR
BP	FRSAC	1.9E-02	2	3.7E-06	0	*	*	2.6E-03	2	7.7E-03	2
	SGCA	*	*	*	*	*	*	*	*	7.4E-03	18
MF	FRSAC	1.7E-03	10	2.1E-05	0	4.3E-02	14	*	*	3.7E-03	4
	SGCA	*	*	4.8E-03	0	*	*	4.8E-04	0	3.0E-03	2
CC	FRSAC	3.2E-03	2	4.5E-02	16	2.8E-05	0	2.8E-02	2	*	*
	SGCA	*	*	*	*	*	*	*	*	*	*

In this context, it should be noted that the performance of the actual gene is better than that of the augmented gene computed from the coarse cluster for breast, leukemia, colon, and prostate cancer data irrespective of the number of clusters and quantitative indexes used. For other data sets, the performance of the augmented gene computed from the coarse cluster is better than the actual gene in only 15 cases out of total 48 cases. Hence, the augmented cluster representatives should be computed by considering only genes of finer clusters, not all genes of corresponding coarse clusters. The attributes present in the coarse cluster, but not in the corresponding finer cluster, reduce the redundancy among the augmented attributes to be generated. The inclusion of these genes with the genes of the finer cluster may degrade the quality of the solution.

D. Biological Significance

To interpret the biological significance of the generated clusters, the gene ontology (GO) term finder is used [25]. It finds the most significantly enriched GO terms associated with the genes belonging to a cluster. The GO project aims to build tree structures and controlled vocabularies, also called ontologies, that describe gene products in terms of their associated biological processes (BPs), molecular functions (MFs), or cellular components (CCs). 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 [25]. The closer the p-value is to zero, the more significant the particular GO term associated with the group of genes becomes, 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 the best cluster produced by the proposed algorithm. The GO term finder is used to compute both the p-value and the FDR (in 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 the best cluster. Table VII presents the p-values and the FDR for the BP, the MF, and the CC on different data sets. The results corresponding to the best clusters of the SGCA [10] are also provided on same data sets for the sake of comparison. The "*" in Table VII represents that no significant shared term is found considering the p-value cutoff as 0.05. From the results reported in Table VII, it is seen that the best cluster generated

TABLE VIII
COMPARATIVE PERFORMANCE ANALYSIS OF DIFFERENT METHODS ON EIGHT MICROARRAY DATA SETS

Microarray Data Sets	Methods / Algorithms	$m = 1$				$m = 2$				$m = 3$			
		SVM	K-NN	NB	CS	SVM	K-NN	NB	CS	SVM	K-NN	NB	CS
Breast ($n = 49$, $D = 7129$)	FRSAC	100	100	100	0.46	100	100	100	0.47	100	100	100	0.45
	FEPM	85.7	89.8	89.8	1.16	91.8	89.8	89.8	1.28	91.8	89.8	89.8	2.05
	RFCM	81.6	83.7	83.7	1.54	81.6	83.7	83.7	1.16	85.7	85.7	83.7	1.29
	SGCA	100	100	100	1.74	100	100	100	1.29	100	100	100	1.83
	FCM	77.6	77.6	83.7	1.67	77.6	81.6	83.7	1.32	81.6	83.7	83.7	1.54
	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
	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 ($n = 72$, $D = 7070$)	FRSAC	100	100	100	0.47	100	100	100	0.47	100	100	100	0.46
	FEPM	94.4	94.4	81.9	0.59	95.8	95.8	95.8	0.99	97.2	97.2	95.8	0.95
	RFCM	79.2	79.2	80.6	1.32	80.6	81.9	80.6	1.17	84.7	84.7	81.9	1.04
	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
	FCM	69.4	73.6	73.6	1.74	77.8	80.6	77.8	2.08	81.9	81.9	80.6	2.24
	ACA	87.4	82.4	88.2	1.69	88.2	87.4	88.2	1.19	88.2	91.2	88.2	3.25
	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 ($n = 62$, $D = 2000$)	FRSAC	100	100	95.2	0.62	100	100	100	0.59	100	100	98.4	0.63
	FEPM	85.5	85.5	83.9	1.52	83.9	85.5	75.8	2.08	85.5	87.1	87.1	2.01
	RFCM	80.6	80.6	82.3	1.65	82.3	82.3	80.6	1.77	83.9	88.7	85.5	1.38
	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
	FCM	72.6	72.6	77.4	2.89	72.6	77.4	77.4	2.10	83.9	82.3	82.3	1.76
	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
	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
Lung ($n = 181$, $D = 12533$)	FRSAC	100	100	98.9	1.28	100	100	100	1.02	100	100	100	0.98
	FEPM	96.7	96.7	96.1	0.55	99.5	99.5	97.2	0.52	100	100	96.7	0.47
	RFCM	84.5	84.5	82.9	1.33	84.5	84.5	83.4	1.27	85.1	85.6	85.1	0.99
	SGCA	96.7	96.7	80.7	1.70	96.7	97.2	96.1	1.89	100	96.7	96.1	1.15
	FCM	87.9	83.4	82.9	3.61	83.4	87.9	87.9	1.94	84.0	83.9	83.9	1.09
	ACA	82.9	82.9	65.2	1.88	82.9	80.7	80.7	1.04	88.4	85.1	82.9	1.03
	mRMR	97.8	97.8	97.8	0.49	96.7	96.7	98.9	0.84	98.9	96.7	98.9	0.84
RBreast ($n = 97$, $D = 24188$)	FRSAC	99.0	100	60.8	0.58	99.0	99.0	89.7	0.58	100	100	95.9	0.71
	FEPM	73.2	72.2	72.2	2.39	73.2	80.4	65.0	3.95	76.3	79.4	78.4	4.08
	RFCM	71.1	72.2	60.8	2.30	72.2	72.2	73.2	1.99	78.4	78.4	75.3	2.16
	SGCA	80.4	80.4	56.7	2.60	80.4	83.5	60.8	3.56	89.7	89.7	78.4	3.18
	FCM	60.8	60.8	55.7	2.87	60.8	60.8	55.7	2.39	71.1	72.2	72.2	2.74
	ACA	53.6	53.6	54.6	2.38	53.6	55.7	60.8	2.70	56.7	54.6	60.8	2.13
	mRMR	71.1	73.2	75.3	4.07	71.1	72.2	80.4	8.46	73.2	71.1	80.4	7.79
Prostate ($n = 136$, $D = 12600$)	FRSAC	100	100	75.0	1.44	100	100	78.7	1.51	100	100	80.2	1.59
	FEPM	74.3	75.7	65.4	4.18	82.4	86.0	91.2	4.18	89.7	86.0	91.2	4.94
	RFCM	80.9	80.9	74.3	2.66	80.9	82.4	78.7	2.73	82.4	82.4	80.2	1.96
	SGCA	82.4	82.4	75.7	2.09	82.4	86.0	78.7	2.17	86.0	86.0	78.7	4.08
	FCM	71.3	74.3	74.3	3.71	71.3	74.3	74.3	3.45	80.9	80.9	80.2	2.83
	ACA	71.3	71.3	50.7	4.11	74.3	80.9	55.1	3.09	74.3	80.9	52.9	2.42
	mRMR	71.3	80.9	84.6	4.19	82.4	86.0	91.2	4.18	88.2	86.8	93.4	4.92
RAOA ($n = 30$, $D = 18432$)	FRSAC	100	100	100	0.52	100	100	100	0.58	100	100	100	0.58
	FEPM	93.3	90.0	93.3	0.87	96.7	96.7	93.3	1.32	100	100	93.3	1.31
	RFCM	86.7	86.7	80.0	1.03	86.7	86.7	80.0	1.66	90.0	86.7	86.7	0.98
	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
	FCM	83.3	83.3	80.0	1.74	83.3	76.7	83.3	1.82	86.7	83.3	86.7	1.16
	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
	mRMR	93.3	90.0	93.3	0.87	96.7	96.7	90.0	1.26	96.7	100	90.0	2.01
RAHC ($n = 50$, $D = 41056$)	FRSAC	100	100	100	0.61	100	100	100	0.57	100	100	100	0.62
	FEPM	78.0	82.0	50.0	1.84	92.0	90.0	92.0	2.55	92.0	92.0	90.0	2.30
	RFCM	68.0	72.0	72.0	1.33	76.0	72.0	72.0	1.06	84.0	88.0	88.0	1.19
	SGCA	92.0	92.0	92.0	1.76	90.0	96.0	92.0	3.08	90.0	96.0	92.0	2.18
	FCM	66.0	66.0	50.0	1.54	70.0	68.0	50.0	1.49	78.0	78.0	80.0	1.22
	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
	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

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

E. Comparative Performance Analysis

Finally, Table VIII compares the performance of the proposed FRSAC algorithm with the best performance of different algorithms, namely, FEPM [21], RFCM [18], SGCA [10], ACA [5], FCM [20], and mRMR [23]. The results are presented based on the best classification accuracy of the SVM, the K-NN rule, and the NB classifier for eight microarray data sets. From the results reported in Table VIII, it is seen that the proposed supervised attribute clustering algorithm generates a set of clusters having the highest classification accuracy of the SVM, the K-NN rule, and the NB classifier and the lowest CS index values in most of the cases. However, the FEPM and mRMR methods perform better than the proposed algorithm for the lung cancer data set with respect to the CS index and for the RBreast data set at $m = 1$ and the prostate cancer data set at $m = 2$ and 3 with respect to the NB classifier. Also, the performance of the proposed algorithm is lesser than that of the SGCA and the mRMR for the prostate cancer data set at $m = 1$ with respect to the NB classifier. That is, the proposed supervised attribute clustering algorithm performs better than the existing algorithms in 89 cases out of a total of 96 cases.

The better performance of the proposed FRSAC algorithm is achieved due to the fact that it uses the fuzzy–rough supervised similarity measure to generate coregulated gene clusters with strong association to the class labels. The measure incorporates the information of sample categories while measuring the similarity between genes. Also, it can deal with uncertainty, vagueness, and incompleteness in the class definition. Moreover, the cluster representatives of the proposed algorithm are modified based on the information of sample categories. 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 sample categories.

V. Conclusion

The main contribution of this paper is threefold, namely, defining a new quantitative measure, based on fuzzy–rough sets, to calculate the similarity between two attributes or genes, which incorporates the information of sample categories; developing a new supervised attribute clustering algorithm to find coregulated clusters of genes whose collective expression is strongly associated with the sample categories; and comparing the performance of the proposed method and some existing methods using the CS index and the predictive accuracy of the SVM, the K-NN rule, and the NB classifier.

For six cancer and two arthritis microarray data sets, significantly better results are found for the proposed method compared with existing methods in most of the cases. All the results reported in this paper demonstrate the feasibility and the effectiveness of the proposed method. It is capable of identifying coregulated clusters of genes whose average expression is strongly associated with the sample categories or class labels. 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, Oct. 1999.

[2] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1982.

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

[4] 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, Nov. 2004.

[5] 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. Comput. Biol. Bioinformatics*, vol. 2, no. 2, pp. 83–101, Apr.–Jun. 2005.

[6] J. Herrero, A. Valencia, and J. Dopazo, "A hierarchical unsupervised growing neural network for clustering gene expression patterns," *Bioinformatics*, vol. 17, no. 2, pp. 126–136, Feb. 2001.

[7] 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, Nov. 1999.

[8] 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. Acad. Sci., U.S.A.*, vol. 96, no. 6, pp. 2907–2912, Mar. 1999.

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

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

[11] T. Hastie, R. Tibshirani, D. Botstein, and P. Brown, "Supervised harvesting of expression trees," *Genome Biol.*, vol. 2, no. 1, pp. 1–12, 2001.

[12] D. Nguyen and D. Rocke, "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39–50, Jan. 2002.

[13] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, pp. 338–353, 1965.

[14] Z. Pawlak, *Rough Sets, Theoretical Aspects of Reasoning About Data*. Dordrecht, The Netherlands: Kluwer, 1991.

[15] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *Int. J. Gen. Syst.*, vol. 17, no. 2/3, pp. 191–209, Jun. 1990.

[16] P. Maji and S. K. Pal, "Feature selection using f-information measures in fuzzy approximation spaces," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 6, pp. 854–867, Jun. 2010.

[17] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approach," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 12, pp. 1457–1471, Dec. 2004.

[18] P. Maji and S. K. Pal, "Rough set based generalized fuzzy C-means algorithm and quantitative indices," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 6, pp. 1529–1540, Dec. 2007.

[19] A. P. Gasch and M. B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy K-means clustering," *Genome Biol.*, vol. 3, no. 11, pp. 1–22, Oct. 2002.

[20] D. Dembele and P. Kastner, "Fuzzy C-means method for clustering microarray data," *Bioinformatics*, vol. 19, no. 8, pp. 973–980, May 2003.

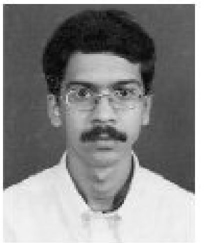
[21] P. Maji and S. K. Pal, "Fuzzy-rough sets for information measures and selection of relevant genes from microarray data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 3, pp. 741–752, Jun. 2010.

[22] S. K. Pal and S. Mitra, *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing*. New York: Wiley, 1999.

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

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

[25] 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, no. 18, pp. 3710–3715, Dec. 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, Kolkata, India, in 1998, 2000, and 2005, respectively.

He is currently an Assistant Professor with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata. He has authored around 60 papers in international journals and conference proceedings. He is also a reviewer of many international journals. His research interests include pattern recognition, computational biology and bioinformatics, medical image processing, cellular automata, and soft computing.

Dr. Maji was a recipient of 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., and the 2009 Young Scientist Award from the National Academy of Sciences, India. He has been selected as the 2009 Associate of the Indian Academy of Sciences, India.