Fuzzy–Rough Sets for Information Measures and Selection of Relevant Genes From Microarray Data

Pradipta Maji and Sankar K. Pal, *Fellow, IEEE*

***Abstract*—Several information measures such as entropy, mutual information, and f-information have been shown to be successful for selecting a set of relevant and nonredundant genes from a high-dimensional microarray data set. However, for continuous gene expression values, it is very difficult to find the true density functions and to perform the integrations required to compute different information measures. In this regard, the concept of the fuzzy equivalence partition matrix is presented to approximate the true marginal and joint distributions of continuous gene expression values. The fuzzy equivalence partition matrix is based on the theory of fuzzy–rough sets, where each row of the matrix represents a fuzzy equivalence partition that can automatically be derived from the given expression values. The performance of the proposed approach is compared with that of existing approaches using the class separability index and the predictive accuracy of the support vector machine. An important finding, however, is that the proposed approach is shown to be effective for selecting relevant and nonredundant continuous-valued genes from microarray data.**

***Index Terms*—Classification, gene selection, information measures, microarray analysis, rough sets.**

I. Introduction

THE 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].

In most gene expression data, the number of training samples is very small compared to the large number of genes involved in the experiments. However, among the large amount of genes, only a small fraction is effective for performing a certain task. Furthermore, a small subset of genes is desirable in developing gene-expression-based diagnostic tools for delivering precise, reliable, and interpretable results. 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 the most relevant genes is the goal of gene selection [1].

In the gene selection process, an optimal gene subset is always relative to a certain criterion. In general, different criteria may lead to different optimal gene subsets. In this regard, several information measures such as entropy, mutual information [2], [3], and f-information [4] have successfully been used in selecting a set of relevant and nonredundant genes from a microarray data set.

However, for real-valued gene expression data, the estimation of different information measures is a difficult task as it requires knowledge on the underlying probability density functions of the data and the integration on these functions. In general, the continuous expression values are divided into several discrete partitions, and the information measures are calculated using the definitions for discrete cases [2], [4]. The inherent error that exists in the discretization process is of major concern in the computation of information measures of continuous gene expression values. In [3] and [5], histograms are used to estimate the true density functions, and the computational difficulty of performing integration can be circumvented in a very efficient way. However, the histogram-based approaches are only applicable to relatively low-dimensional data as the sparse data distribution encountered in a high-dimensional data set may greatly degrade the reliability of histograms [6], [7].

Rough-set theory [8] is a new paradigm to deal with uncertainty, vagueness, and incompleteness. It has been applied to fuzzy rule extraction, reasoning with uncertainty, fuzzy modeling, feature selection, and so forth [9], [10]. It is proposed for indiscernibility in classification according to some similarity [8]. In [11], Hu *et al.* have used the concept of the crisp equivalence relation of rough sets to compute entropy and mutual information in crisp approximation spaces that can be used for feature selection of discrete-valued data sets. However, there are usually real-valued data and fuzzy information in real-world applications. Combining fuzzy sets and rough sets provides an important direction in reasoning with uncertainty for real-valued data sets [9], [12]. Both fuzzy sets and rough sets provide a mathematical framework to capture uncertainties associated with the data [12]. They are complementary in some aspects. The generalized theories of rough–fuzzy sets and fuzzy–rough sets have successfully been applied to feature selection of real-valued data [9], mining stock price, vocabulary for information retrieval, fuzzy decision rule extraction, rough–fuzzy clustering, and so forth [10], [13]. In [11], Hu *et al.* have also used the concept of the fuzzy equivalence relation matrix of fuzzy–rough sets to compute the entropy and mutual information in fuzzy approximation spaces, which can be used for feature selection from real-valued data sets. However, many useful information measures such as several f-information measures cannot be computed from the fuzzy equivalence relation matrix [11] as it does not provide a way to directly compute marginal and joint distributions. Furthermore, the fuzzy–rough-set-based feature selection methods proposed in [9] and [11]

Manuscript received December 1, 2008; revised March 25, 2009 and June 8, 2009. First published November 3, 2009; current version published June 16, 2010. This paper was recommended by Associate Editor L. Wang.

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

Digital Object Identifier 10.1109/TSMCB.2009.2028433

select the relevant or predictive features of a data set without considering the redundancy among them.

In this paper, a new concept of the fuzzy equivalence partition matrix (FEPM) is introduced for computing different information measures on fuzzy approximation spaces. Each row of the matrix represents a fuzzy equivalence partition, which offers an efficient way to estimate the true density functions of continuous-valued gene expression data required for computing different information measures. A subset of genes from the whole gene set is selected by maximizing the relevance and minimizing the redundancy of the selected genes. The relevance and redundancy of the genes are calculated using several information measures on fuzzy approximation spaces based on the concept of the FEPM. The effectiveness of the proposed method, along with a comparison with existing methods, is demonstrated on a set of microarray data sets.

The structure of the rest of this paper is described as follows. Section II briefly introduces the necessary notions of rough sets and fuzzy–rough sets. In Section III, the formulas of Shannon's entropy are introduced for fuzzy approximation spaces with a new concept of the FEPM, along with several information measures. The proposed gene selection method based on information measures for fuzzy approximation spaces is presented in Section IV. A few case studies and a comparison with existing methods are reported in Section V. Concluding remarks are given in Section VI.

II. Rough Sets and Fuzzy–Rough Sets

In this section, the basic notions in the theories of rough sets and fuzzy–rough sets are reported.

A. Rough Sets

The theory of rough sets begins with the notion of an approximation space, which is a pair $\langle \mathbb{U}, \mathbb{A} \rangle$, where $\mathbb{U}$ is a nonempty set (the universe of discourse), i.e., $\mathbb{U} = \{x_1, \ldots, x_i, \ldots, x_n\}$, and $\mathbb{A}$ is a family of attributes, also called knowledge in the universe. V is the value domain of $\mathbb{A}$, and $\hat{f}$ is an information function $\hat{f} : \mathbb{U} \times \mathbb{A} \to V$. An approximation space is also called an information system [8].

Any subset $\mathbb{P}$ of knowledge $\mathbb{A}$ defines an equivalence (also called indiscernibility) relation $IND(\mathbb{P})$ on $\mathbb{U}$

$$IND(\mathbb{P}) = \left\{ (x_i, x_j) \in \mathbb{U} \times \mathbb{U} | \quad \forall a \in \mathbb{P}, \hat{f}_a(x_i) = \hat{f}_a(x_j) \right\}.$$

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 precisely describe X in $\langle \mathbb{U}, \mathbb{A} \rangle$. One may characterize X by a pair of lower and upper approximations defined as follows [8]:

$$\underline{\mathbb{P}}(X) = \bigcup \{[x_i]_{\mathbb{P}} \mid [x_i]_{\mathbb{P}} \subseteq X\}$$

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

That is, 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 is simply called the rough set 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 sometimes called the positive region, denoted as $POS_{\mathbb{P}}(X)$. A set X is said to be definable (or exact) in $\langle \mathbb{U}, \mathbb{A} \rangle$ if and only if $\underline{\mathbb{P}}(X) = \overline{\mathbb{P}}(X)$. Otherwise, X is indefinable and termed as a rough set. $BN_{\mathbb{P}}(X) = \overline{\mathbb{P}}(X) \setminus \underline{\mathbb{P}}(X)$ is called a boundary set.

B. Fuzzy–Rough Sets

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, which are also called fuzzy knowledge granules [9], [12]. 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 be a fuzzy subset of $\mathbb{U}$. The fuzzy $\mathbb{P}$-lower and $\mathbb{P}$-upper approximations are then defined as follows [12]:

$$\mu_{\underline{\mathbb{P}}X}(F_i) = \inf_x \{\max \{(1 - \mu_{F_i}(x)), \mu_X(x)\}\}, \quad \forall I \tag{2}$$

$$\mu_{\overline{\mathbb{P}}X}(F_i) = \sup_x \{\min \{\mu_{F_i}(x), \mu_x(x)\}\}, \quad \forall i \tag{3}$$

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. Note that, although the universe of discourse in feature selection is finite, this is not the case in general, hence the use of sup and inf. 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 of this, the fuzzy lower and upper approximations can be defined as [9]

$$\mu_{\underline{\mathbb{P}}X}(x) = \sup_{F_i \in \mathbb{U}/\mathbb{P}} \min \left\{ \mu_{F_i}(x), \mu_{\underline{\mathbb{P}}X}(F_i) \right\} \tag{4}$$

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

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.

III. FEPM

Shannon's information entropy works in the case where a crisp equivalence relation or a partition is defined, that is, it is suitable for Pawlak's approximation space [11]. In this section, a new formula to compute Shannon's entropy with a FEPM is presented that will be used to measure information on fuzzy approximation spaces.

A. Entropy on Fuzzy Approximation Spaces

Given a finite set $\mathbb{U}$, $\mathbb{A}$ is a fuzzy 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 the c partitions of $\mathbb{U}$ are sets of (cn) values $\{m_{ij}^{\mathbb{A}}\}$ that can conveniently be arrayed as a $(c \times n)$ matrix $\mathbb{M}_{\mathbb{A}} = [m_{ij}^{\mathbb{A}}]$. The matrix $\mathbb{M}_{\mathbb{A}}$ is termed as the FEPM and is denoted by

$$\mathbb{M}_{\mathbb{A}} = \begin{pmatrix} m_{11}^{\mathbb{A}} & m_{12}^{\mathbb{A}} & \cdots & m_{1n}^{\mathbb{A}} \\ m_{21}^{\mathbb{A}} & m_{22}^{\mathbb{A}} & \cdots & m_{2n}^{\mathbb{A}} \\ \cdots & \cdots & \cdots & \cdots \\ m_{c1}^{\mathbb{A}} & m_{c2}^{\mathbb{A}} & \cdots & m_{cn}^{\mathbb{A}} \end{pmatrix} \tag{6}$$

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

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

where $m_{ij}^{\mathbb{A}} \in [0, 1]$ represents the membership of object x_j in the ith fuzzy equivalence partition or class F_i. The aforementioned axioms should hold for every fuzzy equivalence partition, which correspond to the requirement that an equivalence class is nonempty. Obviously, this definition degenerates to the normal definition of equivalence classes when the equivalence relation is nonfuzzy.

A $c \times n$ FEPM $\mathbb{M}_{\mathbb{A}}$ represents the c fuzzy equivalence partitions of the universe generated by a fuzzy equivalence relation. Each row of the matrix $\mathbb{M}_{\mathbb{A}}$ is a fuzzy equivalence partition or class. The ith fuzzy equivalence partition is therefore given by

$$F_i = \left\{ m_{i1}^{\mathbb{A}}/x_1 + m_{i2}^{\mathbb{A}}/x_2 + \cdots + m_{in}^{\mathbb{A}}/x_n \right\}. \tag{7}$$

As to a fuzzy partition induced by a fuzzy equivalence relation, the equivalence class is a fuzzy set. "+" means the operator of union in this case. The cardinality of the fuzzy set F_i can be calculated with

$$|F_i| = \sum_{j=1}^{n} m_{ij}^{\mathbb{A}} \tag{8}$$

which appears to be a natural generalization of the crisp set. The information quantity of a fuzzy attribute set $\mathbb{A}$ or fuzzy equivalence partition is then defined as

$$H(\mathbb{A}) = -\sum_{i=1}^{c} \lambda_{F_i} \log \lambda_{F_i} \tag{9}$$

where $\lambda_{F_i} = (|F_i|/n)$, called a fuzzy relative frequency, and c is the number of fuzzy equivalence partitions or classes. The measure $H(\mathbb{A})$ has the same form as the Shannon's entropy. The information quantity or the entropy value monotonously increases with the discernibility power of the fuzzy attributes. Combining (8) and (9), the form of Shannon's entropy, in terms of the FEPM, on fuzzy approximation spaces is given by

$$H(\mathbb{A}) = -\sum_{i=1}^{c} \left[\frac{1}{n} \sum_{k=1}^{n} m_{ik}^{\mathbb{A}} \right] \log \left[\frac{1}{n} \sum_{k=1}^{n} m_{ik}^{\mathbb{A}} \right]. \tag{10}$$

B. Mutual Information on Fuzzy Approximation Spaces

Given $\langle \mathbb{U}, \mathbb{A} \rangle$, $\mathbb{P}$ and $\mathbb{Q}$ are two subsets of $\mathbb{A}$. The information quantity corresponding to $\mathbb{P}$ and $\mathbb{Q}$ are given by

$$H(\mathbb{P}) = -\sum_{i=1}^{p} \lambda_{P_i} \log \lambda_{P_i} \quad H(\mathbb{Q}) = -\sum_{j=1}^{q} \lambda_{Q_j} \log \lambda_{Q_j}$$

where p and q are the numbers of fuzzy equivalence partitions or classes generated by the fuzzy attribute sets $\mathbb{P}$ and $\mathbb{Q}$, respectively, and P_i and Q_j represent the corresponding ith and jth fuzzy equivalence partitions. The joint entropy of $\mathbb{P}$ and $\mathbb{Q}$ can be defined as follows:

$$H(\mathbb{PQ}) = -\sum_{k=1}^{r} \lambda_{R_k} \log \lambda_{R_k} \tag{11}$$

where r is the number of resultant fuzzy equivalence partitions, R_k is the corresponding kth equivalence partition, and λ_{R_k} is the joint frequency of P_i and Q_j, which is given by

$$\lambda_{R_k} = \lambda_{P_i Q_j} = \frac{|P_i \cap Q_j|}{n}, \qquad \text{where } k = (i-1)q + j. \tag{12}$$

In other words, the joint frequency λ_{R_k} can be calculated from the $r \times n$ FEPM $\mathbb{M}_{\mathbb{PQ}}$, where

$$\mathbb{M}_{\mathbb{PQ}} = \mathbb{M}_{\mathbb{P}} \cap \mathbb{M}_{\mathbb{Q}} \quad m_{kl}^{\mathbb{PQ}} = m_{il}^{\mathbb{P}} \cap m_{jl}^{\mathbb{Q}}. \tag{13}$$

Hence, the mutual information between two fuzzy attribute sets $\mathbb{P}$ and $\mathbb{Q}$ is given by

$$I(\mathbb{PQ}) = H(\mathbb{P}) + H(\mathbb{Q}) - H(\mathbb{PQ}). \tag{14}$$

Combining (8), (13), and (14), the mutual information between two fuzzy attribute sets $\mathbb{P}$ and $\mathbb{Q}$, in terms of the FEPM, can be represented as

$$\begin{aligned} I(\mathbb{P}, \mathbb{Q}) = & -\sum_{i=1}^{p} \left[\frac{1}{n} \sum_{k=1}^{n} m_{ik}^{\mathbb{P}} \right] \log \left[\frac{1}{n} \sum_{k=1}^{n} m_{ik}^{\mathbb{P}} \right] \\ & -\sum_{j=1}^{q} \left[\frac{1}{n} \sum_{k=1}^{n} m_{jk}^{\mathbb{Q}} \right] \log \left[\frac{1}{n} \sum_{k=1}^{n} m_{jk}^{\mathbb{Q}} \right] \\ & +\sum_{i=1}^{p} \sum_{j=1}^{q} \left[\frac{1}{n} \sum_{k=1}^{n} \left(m_{ik}^{\mathbb{P}} \cap m_{jk}^{\mathbb{Q}} \right) \right] \\ & \times \log \left[\frac{1}{n} \sum_{k=1}^{n} \left(m_{ik}^{\mathbb{P}} \cap m_{jk}^{\mathbb{Q}} \right) \right]. \end{aligned} \tag{15}$$

C. Other Information Measures

In this paper, two frequently used information measures, i.e., V-information and χ^2-information, are also reported for fuzzy approximation spaces based on the concept of the FEPM.

1) V-Information: In fuzzy approximation spaces, one of the simplest measures of dependence can be obtained using the function $V = |x - 1|$, which results in the V-information

$$V(R\|P \times Q) = \sum_{i,j,k} |\lambda_{R_k} - \lambda_{P_i}\lambda_{Q_j}| \tag{16}$$

where $P = \{\lambda_{P_i} | i = 1, 2, \ldots, p\}$, $Q = \{\lambda_{Q_j} | j = 1, 2, \ldots, q\}$, and $R = \{\lambda_{R_k} | k = 1, 2, \ldots, r\}$ represent two marginal frequency distributions and their joint frequency distribution, respectively. That is, the V-information calculates the absolute distance between the joint frequency of two fuzzy variables and their marginal frequencies' product.

Combining (8), (13), and (16), the V-information between two fuzzy attribute sets $\mathbb{P}$ and $\mathbb{Q}$, in terms of the FEPM, can be represented by

$$V(\mathbb{P}, \mathbb{Q}) = \sum_{i,j} \left| \frac{1}{n} \sum_{k=1}^{n} \left(m_{ik}^{\mathbb{P}} \cap m_{jk}^{\mathbb{Q}} \right) - \frac{1}{n^2} \sum_{k=1}^{n} m_{ik}^{\mathbb{P}} \sum_{k=1}^{n} m_{jk}^{\mathbb{Q}} \right|.$$

2) χ^2-Information: The χ^2-information measure on fuzzy approximation spaces can be defined as follows:

$$\chi^2(R\|P \times Q) = \sum_{i,j,k} \frac{|\lambda_{R_k} - \lambda_{P_i}\lambda_{Q_j}|^2}{(\lambda_{P_i}\lambda_{Q_j})}. \tag{17}$$

Combining (8), (13), and (17), the χ^2-information between two fuzzy attribute sets $\mathbb{P}$ and $\mathbb{Q}$, in terms of the FEPM, can be represented by

$$\chi^2(\mathbb{P}, \mathbb{Q}) = \sum_{i,j} \frac{\left| \frac{1}{n} \sum_{k=1}^{n} \left(m_{ik}^{\mathbb{P}} \cap m_{jk}^{\mathbb{Q}} \right) - \frac{1}{n^2} \sum_{k=1}^{n} m_{ik}^{\mathbb{P}} \sum_{k=1}^{n} m_{jk}^{\mathbb{Q}} \right|^2}{\frac{1}{n^2} \sum_{k=1}^{n} m_{ik}^{\mathbb{P}} \sum_{k=1}^{n} m_{jk}^{\mathbb{Q}}}.$$

3) Comment: In this context, it should be noted that Hu *et al.* have introduced the concept of the $n \times n$ fuzzy equivalence relation matrix in [11] to compute the entropy and mutual information in fuzzy approximation spaces. However, this matrix does not provide a way to directly compute marginal and joint distributions. In effect, many useful information measures cannot be computed from the fuzzy equivalence relation matrix [11]. Furthermore, the complexity of this approach is $\mathcal{O}(n^2)$, which is higher than $\mathcal{O}(cn)$ of the proposed approach as $c \ll n$.

IV. Proposed Gene Selection Method

In microarray data analysis, the data set may contain a number of redundant genes with low relevance to the classes. The presence of such redundant and nonrelevant genes leads to a reduction in the useful information. Ideally, the selected genes should have high relevance with the classes while the redundancy among them would be as low as possible. The genes with high relevance are expected to be able to predict the classes of the samples. However, the prediction capability is reduced if many redundant genes are selected. In contrast, a microarray data set that contains genes that are not only with high relevance with respect to the classes but also with low mutual redundancy is more effective in its prediction capability. Hence, to assess the effectiveness of the genes, both relevance and redundancy need to be quantitatively measured. An information-measure-based gene selection method is presented here to address this problem.

A. Gene Selection Using Information Measures

Let $\mathbb{G} = \{\mathbb{G}_1, \ldots, \mathbb{G}_i, \ldots, \mathbb{G}_j, \ldots, \mathbb{G}_d\}$ denote the set of genes or fuzzy condition attributes of a given microarray data set and $\mathbb{S}$ be the set of selected genes. Define $\tilde{f}(\mathbb{G}_i, \mathbb{D})$ as the relevance of gene $\mathbb{G}_i$ (fuzzy condition attribute) with respect to class $\mathbb{D}$ (fuzzy decision attribute) and $\tilde{f}(\mathbb{G}_i, \mathbb{G}_j)$ as the redundancy between two genes $\mathbb{G}_i$ and $\mathbb{G}_j$ (fuzzy condition attributes). The total relevance of all selected genes is therefore given by

$$\mathcal{J}_{\text{relev}} = \sum_{\mathbb{G}_i \in \mathbb{S}} \tilde{f}(\mathbb{G}_i, \mathbb{D}) \tag{18}$$

while the total redundancy among the selected genes is

$$\mathcal{J}_{\text{redun}} = \sum_{\mathbb{G}_i, \mathbb{G}_j \in \mathbb{S}} \tilde{f}(\mathbb{G}_i, \mathbb{G}_j). \tag{19}$$

Therefore, the problem of selecting a set $\mathbb{S}$ of nonredundant and relevant genes from the whole set of genes $\mathbb{G}$ (condition attributes) is equivalent to maximizing $\mathcal{J}_{\text{relev}}$ and minimizing $\mathcal{J}_{\text{redun}}$, that is, to maximize the objective function $\mathcal{J}$, where

$$\mathcal{J} = \mathcal{J}_{\text{relev}} - \mathcal{J}_{\text{redun}} = \sum_{i} \tilde{f}(\mathbb{G}_i, \mathbb{D}) - \sum_{i,j} \tilde{f}(\mathbb{G}_i, \mathbb{G}_j). \tag{20}$$

The aforementioned gene selection problem is solved using the following greedy algorithm:

1) Initialize $\mathbb{G} \leftarrow \{\mathbb{G}_1, \ldots, \mathbb{G}_i, \ldots, \mathbb{G}_j, \ldots, \mathbb{G}_d\}$, $\mathbb{S} \leftarrow \emptyset$.
2) Generate FEPM $\mathbb{M}_{\mathbb{G}_i}$ for each gene $\mathbb{G}_i \in \mathbb{G}$.
3) Calculate the relevance $\tilde{f}(\mathbb{G}_i, \mathbb{D})$ of each gene $\mathbb{G}_i \in \mathbb{G}$.
4) Select gene $\mathbb{G}_i$ as the first gene that has the highest relevance $\tilde{f}(\mathbb{G}_i, \mathbb{D})$. In effect, $\mathbb{G}_i \in \mathbb{S}$, and $\mathbb{G} = \mathbb{G} \setminus \mathbb{G}_i$.
5) Generate resultant FEPM $\mathbb{M}_{\mathbb{G}_i\mathbb{G}_j}$ between selected gene $\mathbb{G}_i$ of $\mathbb{S}$ and each of the remaining genes $\mathbb{G}_j$ of $\mathbb{G}$.
6) Calculate the redundancy $\tilde{f}(\mathbb{G}_i, \mathbb{G}_j)$ between selected genes of $\mathbb{S}$ and each of the remaining genes of $\mathbb{G}$.
7) From the remaining genes of $\mathbb{G}$, select gene $\mathbb{G}_j$ that maximizes

$$\tilde{f}(\mathbb{G}_j, \mathbb{D}) - \frac{1}{|\mathbb{S}|} \sum_{\mathbb{G}_i \in \mathbb{S}} \tilde{f}(\mathbb{G}_i, \mathbb{G}_j).$$

   As a result of that, $\mathbb{G}_j \in \mathbb{S}$, and $\mathbb{G} = \mathbb{G} \setminus \mathbb{G}_j$.
8) Repeat the aforementioned three steps until the desired number of genes are selected.

Both the relevance $\tilde{f}(\mathbb{G}_i, \mathbb{D})$ of a gene $\mathbb{G}_i$ (fuzzy condition attribute) with respect to the class labels $\mathbb{D}$ (fuzzy decision attribute) and the redundancy $\tilde{f}(\mathbb{G}_i, \mathbb{G}_j)$ between two genes $\mathbb{G}_i$ and $\mathbb{G}_j$ (fuzzy condition attributes) can be calculated using any one of the information measures on fuzzy approximation spaces reported earlier.

B. Fuzzy Equivalence Classes

In the proposed gene selection 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 $\pi(x; \bar{c}, \sigma)$ [14] represents a set of points clustered around $\bar{c}$, where

$$\pi(x;\bar{c},\sigma) = \begin{cases} 2\left(1-\frac{\|x-\bar{c}\|}{\sigma}\right)^2, & \text{for } \frac{\sigma}{2} \le \|x-\bar{c}\| \le \sigma \\ 1-2\left(\frac{\|x-\bar{c}\|}{\sigma}\right)^2, & \text{for } 0 \le \|x-\bar{c}\| \le \frac{\sigma}{2} \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

where $\sigma > 0$ is the radius of the π function with $\bar{c}$ as the central point, and $\|\cdot\|$ denotes the Euclidean norm. When 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 [14].

Each input real-valued gene in 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 radii of the π functions along each gene axis are automatically determined from the distribution of the training patterns.

1) Computation of Parameters of the π Function: Parameters $\bar{c}$ and σ of each π fuzzy set are computed according to the following procedure [14]. Let $\bar{m}_i$ be the mean of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$ along the ith gene $\mathbb{G}_i$. 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 $[\mathbb{G}_{i_{\min}}, \bar{m}_i)$ and $(\bar{m}_i, \mathbb{G}_{i_{\max}}]$, respectively, where $\mathbb{G}_{i_{\max}}$ and $\mathbb{G}_{i_{\min}}$ denote the upper and lower bounds of the dynamic range of gene $\mathbb{G}_i$ for the training set. For three fuzzy sets (low, medium, and high), the centers and corresponding radii are defined as follows:

$$\bar{c}_{\text{low}}(\mathbb{G}_i) = \bar{m}_{i_l} \quad \bar{c}_{\text{medium}}(\mathbb{G}_i) = \bar{m}_i \quad \bar{c}_{\text{high}}(\mathbb{G}_i) = \bar{m}_{i_h} \tag{22}$$

$$\begin{aligned} \sigma_{\text{low}}(\mathbb{G}_i) &= 2\left(\bar{c}_{\text{medium}}(\mathbb{G}_i) - \bar{c}_{\text{low}}(\mathbb{G}_i)\right) \\ \sigma_{\text{high}}(\mathbb{G}_i) &= 2(\bar{c}_{\text{high}}(\mathbb{G}_i) - \bar{c}_{\text{medium}}(\mathbb{G}_i)) \\ \sigma_{\text{medium}}(\mathbb{G}_i) &= \eta \times \frac{\text{A}}{\text{B}} \end{aligned} \tag{23}$$

where

$$\begin{aligned} \text{A} &= \{\sigma_{\text{low}}(\mathbb{G}_i)\left(\mathbb{G}_{i_{\max}} - c_{\text{medium}}(\mathbb{G}_i)\right) + \sigma_{\text{high}}(\mathbb{G}_i) \\ &\quad \times \left(c_{\text{medium}}(\mathbb{G}_i) - \mathbb{G}_{i_{\min}}\right)\}; \\ \text{B} &= \{\mathbb{G}_{i_{\max}} - \mathbb{G}_{i_{\min}}\} \end{aligned}$$

where η is a multiplicative parameter controlling the extent of 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. Furthermore, the amount of overlap between three fuzzy sets can be different along a different axis, depending on the distribution of patterns.

2) Generation of the FEPM: The $c \times n$ FEPM $\mathbb{M}_{\mathbb{G}_i}$, corresponding to the ith gene $\mathbb{G}_i$, can be calculated from the c fuzzy equivalence classes of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$, where

$$m_{kj}^{\mathbb{G}_i} = \frac{\pi(x_j; \bar{c}_k, \sigma_k)}{\sum_{l=1}^{c} \pi(x_j; \bar{c}_l, \sigma_l)}. \tag{24}$$

Corresponding to three fuzzy sets, i.e., low, medium, and high ($c = 3$), the following relations hold:

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

In effect, each position $m_{kj}^{\mathbb{G}_i}$ of the FEPM $\mathbb{M}_{\mathbb{G}_i}$ must satisfy the following conditions:

$$m_{kj}^{\mathbb{G}_i} \in [0,1]; \sum_{k=1}^{c} m_{kj}^{\mathbb{G}_i} = 1, \forall j, \text{ and for any value of } k, \text{ if}$$

$$s = \arg\max_j \left\{m_{kj}^{\mathbb{G}_i}\right\}, \text{ then } \max_j \left\{m_{kj}^{\mathbb{G}_i}\right\} = \max_l \left\{m_{ls}^{\mathbb{G}_i}\right\} > 0.$$

V. Experimental Results and Discussion

The performance of the proposed FEPM-based density approximation approach is extensively compared with that of two existing methods: 1) the discretization-based approach (discrete) [2], [4] and 2) the Parzen-window-based approach (Parzen) [5]. Results are reported with respect to three widely used information measures, i.e., mutual information, V-information, and χ^2-information. All these measures are applied to calculate both gene–class relevance and gene–gene redundancy. To analyze the performance of the proposed and existing methods, the experimentation is done on five microarray gene expression data sets. The major metrics for evaluating the performance of different methods are the class separability index [15] and the classification accuracy of the support vector machine (SVM) [16].

A. Gene Expression Data Sets

In this paper, three publicly available cancer and two publicly available arthritis data sets are used. Since binary classification is a typical and fundamental issue in the 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 7129 genes in 49 breast tumor samples [17]. The samples are classified according to their estrogen receptor (ER) status: 25 samples are ER positive, while the other 24 samples are ER negative.

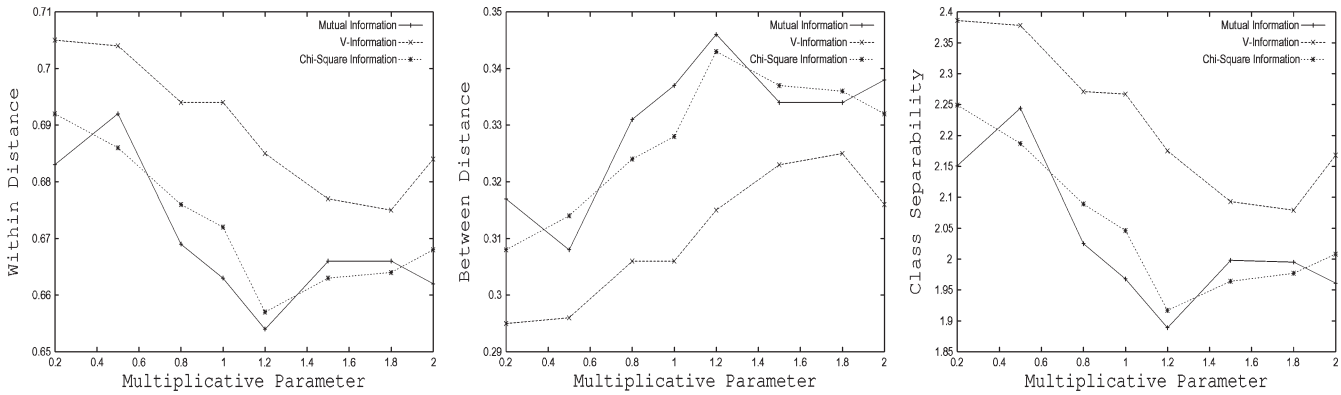


Fig. 1. Variation of the class separability index with respect to multiplicative parameter η for the breast cancer data set. (a) Within-class scatter matrix. (b) Between-class scatter matrix. (c) Class separability index.

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

3) Colon Cancer: The colon cancer data set contains expression levels of 40 tumor and 22 normal colon tissues. Only the 2000 genes with the highest minimal intensity were selected in [18].

4) RAOA: The rheumatoid arthritis versus osteoarthritis (RAOA) data set consists of gene expression profiles of 30 patients: 21 with rheumatoid arthritis (RA) and 9 with osteoarthritis [19]. The Cy5-labeled experimental cDNA and the Cy3-labeled common reference sample were pooled and hybridized to the lymphochips containing ∼18 000 cDNA spots representing genes of relevance in immunology [19].

5) RAHC: The RA versus healthy controls (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 sex-matched healthy controls performed on microarrays with a complexity of ∼26 000 unique genes (43 000 elements) [20].

B. Class Prediction Methods

The following two quantitative indices are used to evaluate the performance of different methods.

1) Class Separability Index: The class separability index $\mathcal{S}$ [15] of a data set is defined as $\mathcal{S} = \text{trace}(S_b^{-1} S_w)$, where S_w is the within-class scatter matrix, and S_b is the between-class scatter matrix, defined as follows:

$$S_w = \sum_{j=1}^{C} p_j E\left\{(X-\mu_j)(X-\mu_j)^T | c_j\right\} = \sum_{j=1}^{C} p_j \Sigma_j$$

$$S_b = \sum_{j=1}^{C} (\mu_j - M_0)(\mu_j - M_0)^T, \qquad M_0 = E\{X\} = \sum_{j=1}^{C} p_j \mu_j$$

where C is the number of classes, p_j is the *a priori* probability that a pattern belongs to class c_j, X is a feature vector, M_0 is the sample mean vector for the entire data points, μ_j is the sample mean vector of class c_j, Σ_j is the sample covariance matrix of class c_j, and $E\{\cdot\}$ is the expectation operator. A lower value of the separability criteria ensures that the classes are well separated by their scatter means.

2) SVM: The SVM [16] is a 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.

C. Performance Analysis of the FEPM

η is a multiplicative parameter controlling the extent of overlapping between low and medium fuzzy sets or medium and high fuzzy sets. Keeping the values of σ_{low} and σ_{high} fixed, the amount of overlapping among the three π functions can be altered by varying σ_{medium}. As η is decreased, the radius σ_{medium} decreases around $\bar{c}_{\text{medium}}$ such that, ultimately, there is insignificant overlapping between the low and medium π functions or medium and high π functions. This implies that certain regions along the ith gene axis $\mathbb{G}_i$ go underrepresented such that the three membership values corresponding to the three fuzzy sets, i.e., low, medium, and high, attain small values. Note that the particular choice of the values of σs and $\bar{c}$s ensure that, for any pattern x_j along the ith gene axis $\mathbb{G}_i$, at least one of the membership values should be greater than 0.5. On the other hand, as η is increased, the radius σ_{medium} increases around $\bar{c}_{\text{medium}}$ such that the amount of overlapping between the π functions increases.

1) Class Separability Analysis: Figs. 1–5 depict the performance of the proposed method for five microarray data sets in terms of the within-class scatter matrix, between-class scatter matrix, and class separability index. Results are presented for 30 top-ranked genes selected by the proposed method for three information measures and five microarray data sets. Each data set is preprocessed by standardizing each sample to zero mean and unit variance. From the results reported in Figs. 1–5, it can be seen that, as the value of multiplicative parameter η increases, the values of the within-class scatter matrix and class separability index decrease, while the between-class scatter matrix increases, irrespective of the data sets and information

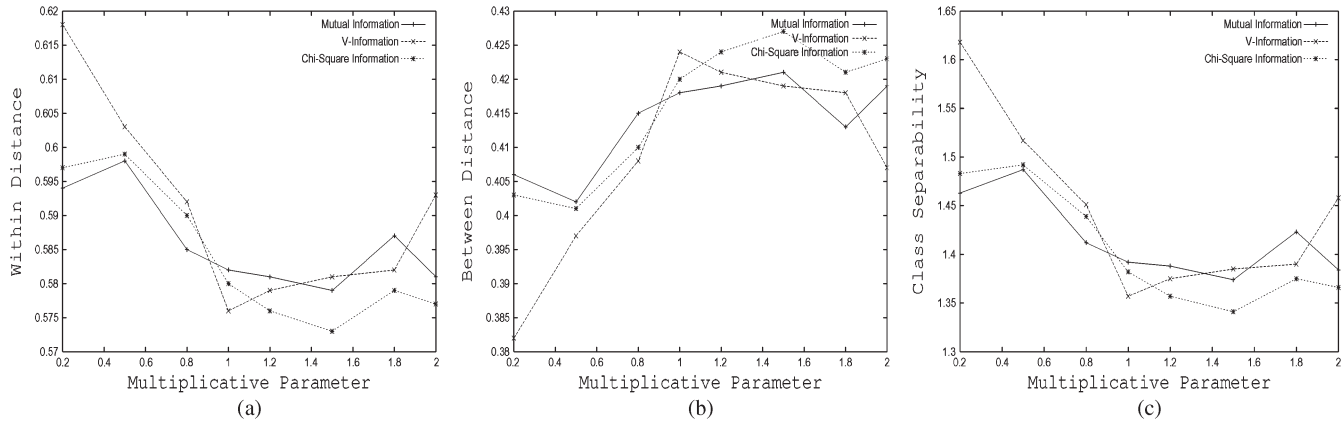


Fig. 2. Variation of the class separability index with respect to multiplicative parameter η for the leukemia data set. (a) Within-class scatter matrix. (b) Between-class scatter matrix. (c) Class separability index.

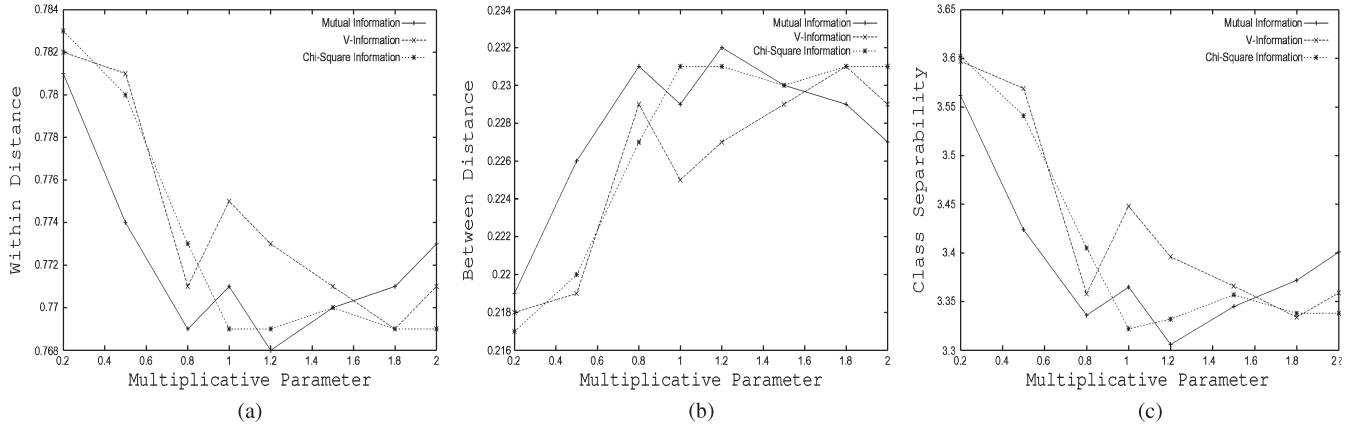


Fig. 3. Variation of the class separability index with respect to multiplicative parameter η for the colon cancer data set. (a) Within-class scatter matrix. (b) Between-class scatter matrix. (c) Class separability index.

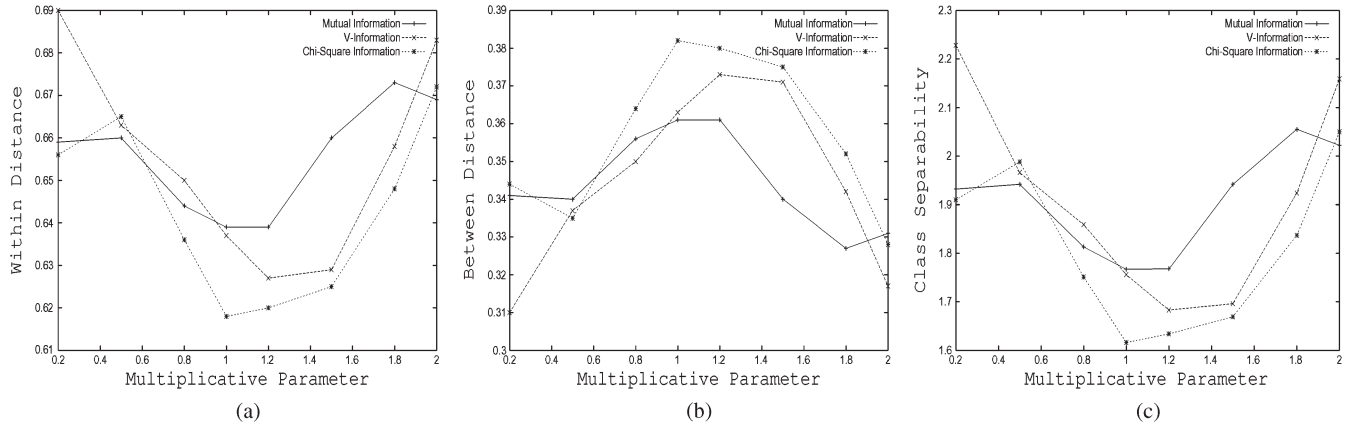


Fig. 4. Variation of the class separability index with respect to multiplicative parameter η for the RAOA data set. (a) Within-class scatter matrix. (b) Between-class scatter matrix. (c) Class separability index.

measures used. The best performance of the proposed method is achieved for $1.0 \leq \eta \leq 1.8$. For $\eta > 1.8$, the performance of the proposed method decreases with the increase in η.

Table I presents the best performance achieved by the proposed method for different data sets and information measures used in terms of the within-class scatter matrix (S_w), between-class scatter matrix (S_b), and class separability index $(\mathcal{S})$, along with the corresponding η value. The proposed method achieves best performance with $\eta = 1.0$ for leukemia data using V-information, colon cancer data using χ^2-information, RAOA data using mutual information and χ^2-information, and RAHC data using mutual information, respectively. Similarly, the best performance is achieved with $\eta = 1.2$ for breast cancer data using mutual information and χ^2-information, colon cancer data using mutual information, RAOA data using V-information, and RAHC data using χ^2-information, respectively. At $\eta = 1.5$,

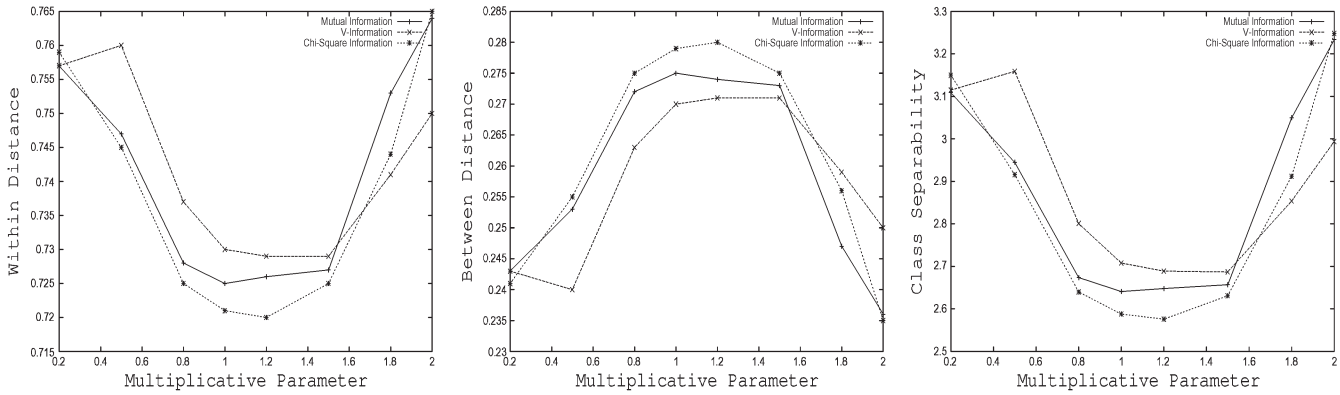


Fig. 5. Variation of the class separability index with respect to multiplicative parameter η for the RAHC data set. (a) Within-class scatter matrix. (b) Between-class scatter matrix. (c) Class separability index.

TABLE I
CLASS SEPARABILITY ANALYSIS

Data Set	Measure	η	S_w	S_b	$\mathcal{S}$
Breast	I	1.2	0.654	0.346	1.889
	V	1.5	0.675	0.325	2.079
	χ^2	1.2	0.657	0.343	1.917
Leukemia	I	1.5	0.579	0.421	1.374
	V	1.0	0.576	0.424	1.357
	χ^2	1.5	0.573	0.427	1.341
Colon	I	1.2	0.768	0.232	3.306
	V	1.8	0.769	0.231	3.334
	χ^2	1.0	0.769	0.231	3.322
RAOA	I	1.0	0.639	0.361	1.767
	V	1.2	0.627	0.373	1.683
	χ^2	1.0	0.618	0.382	1.616
RAHC	I	1.0	0.725	0.275	2.641
	V	1.5	0.729	0.271	2.687
	χ^2	1.2	0.720	0.280	2.576

mutual information and χ^2-information provide the best result for leukemia data, while V-information gives the best performance for RAHC data. On the other hand, V-information provides the best performance for both breast and colon cancer data with $\eta = 1.8$. However, for $\eta > 1.8$, the performance of the proposed method decreases with the increase in η for the three measures and five data sets used.

2) Classification Accuracy Analysis: Tables II–XVI represent the performance of the proposed method in terms of the classification accuracy of the SVM for different values of η. Results are presented for five microarray data sets considering three widely used information measures, i.e., mutual information, V-information, and χ^2-information. To compute the prediction accuracy of the SVM, the leave-one-out cross validation is performed on each gene expression data set. The values of η investigated are 0.2, 0.5, 0.8, 1.0, 1.2, 1.5, 1.8, and 2.0. The number of genes selected ranges from 1 to 30; however, results are reported only for 20 top-ranked genes, and each data set is preprocessed by standardizing each sample to zero mean and unit variance.

Tables II–IV depict the results for the breast cancer data set with respect to three information measures. The 100% classification accuracy of the SVM is obtained for mutual information and χ^2-information considering eight and seven top-ranked genes, respectively, with both $\eta = 0.5$ and 0.8, while

TABLE II
PERFORMANCE OF MUTUAL INFORMATION ON BREAST CANCER DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	85.7	85.7	85.7	85.7	85.7	85.7	85.7	85.7
2	89.8	89.8	89.8	89.8	89.8	91.8	91.8	95.9
3	91.8	91.8	91.8	91.8	91.8	91.8	91.8	93.9
4	87.8	95.9	95.9	95.9	95.9	91.8	91.8	93.9
5	89.8	95.9	95.9	95.9	95.9	87.8	93.9	93.9
6	87.8	93.9	93.9	93.9	93.9	83.7	91.8	91.8
7	93.9	93.9	91.8	93.9	93.9	85.7	93.9	91.8
8	93.9	100	100	91.8	93.9	91.8	89.8	89.8
9	93.9	98.0	100	91.8	91.8	91.8	85.7	89.8
10	91.8	100	100	87.8	93.9	91.8	85.7	85.7
11	89.8	100	100	87.8	89.8	91.8	87.8	85.7
12	91.8	98.0	95.9	89.8	89.8	91.8	87.8	85.7
13	89.8	98.0	98.0	95.9	87.8	93.9	91.8	87.8
14	93.9	98.0	95.9	100	87.8	93.9	91.8	87.8
15	93.9	98.0	95.9	100	87.8	95.9	87.8	87.8
16	93.9	93.9	98.0	100	89.8	95.9	87.8	87.8
17	93.9	95.9	98.0	100	87.8	89.8	87.8	87.8
18	89.8	93.9	100	100	89.8	89.8	91.8	87.8
19	93.9	93.9	98.0	100	100	95.9	95.9	87.8
20	93.9	100	100	100	100	93.9	95.9	91.8

TABLE III
PERFORMANCE OF V-INFORMATION ON BREAST CANCER DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	85.7	85.7	85.7	85.7	85.7	85.7	85.7	85.7
2	93.9	93.9	93.9	89.8	93.9	93.9	93.9	93.9
3	93.9	93.9	93.9	91.8	93.9	95.9	93.9	93.9
4	93.9	98.0	93.9	95.9	93.9	93.9	91.8	95.9
5	93.9	100	93.9	93.9	93.9	95.9	95.9	95.9
6	98.0	100	87.8	98.0	93.9	95.9	91.8	91.8
7	98.0	100	89.8	91.8	93.9	95.9	85.7	85.7
8	95.9	100	95.9	93.9	93.9	91.8	93.9	85.7
9	98.0	100	95.9	93.9	95.9	91.8	95.9	87.8
10	93.9	100	95.9	100	100	95.9	95.9	87.8
11	95.9	95.9	93.9	98.0	100	95.9	95.9	95.9
12	93.9	95.9	91.8	100	100	95.9	95.9	95.9
13	93.9	95.9	91.8	98.0	100	95.9	93.9	93.9
14	95.9	95.9	89.8	98.0	98.0	93.9	93.9	93.9
15	95.9	93.9	91.8	98.0	98.0	93.9	93.9	91.8
16	95.9	95.9	91.8	95.9	98.0	93.9	98.0	98.0
17	98.0	95.9	95.9	95.9	98.0	95.9	98.0	98.0
18	98.0	95.9	95.9	98.0	98.0	95.9	95.9	95.9
19	98.0	95.9	95.9	98.0	98.0	95.9	95.9	95.9
20	98.0	95.9	98.0	95.9	98.0	95.9	95.9	93.9

in case of V-information, five top-ranked genes are required to achieve this accuracy with $\eta = 0.5$. Similarly, a maximum of 98.6% accuracy in the case of leukemia data is obtained for both mutual information and χ^2-information using nine genes

TABLE IV
PERFORMANCE OF χ^2-INFORMATION ON BREAST CANCER DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	85.7	85.7	85.7	85.7	85.7	85.7	85.7	85.7
2	89.8	89.8	89.8	89.8	89.8	91.8	91.8	91.8
3	91.8	91.8	93.9	91.8	91.8	91.8	91.8	89.8
4	95.9	95.9	95.9	95.9	95.9	91.8	91.8	89.8
5	95.9	95.9	95.9	95.9	95.9	87.8	87.8	87.8
6	93.9	93.9	93.9	93.9	93.9	93.9	87.8	87.8
7	93.9	100	100	93.9	95.9	93.9	91.8	91.8
8	91.8	98.0	100	89.8	93.9	93.9	91.8	91.8
9	93.9	100	100	93.9	91.8	93.9	91.8	89.8
10	93.9	100	98.0	91.8	89.8	93.9	89.8	85.7
11	93.9	100	100	95.9	89.8	89.8	87.8	83.7
12	91.8	98.0	100	100	89.8	87.8	87.8	83.7
13	87.8	98.0	100	98.0	91.8	85.7	87.8	85.7
14	95.9	95.9	100	100	89.8	83.7	85.7	95.9
15	95.9	98.0	100	100	89.8	93.9	95.9	95.9
16	98.0	91.8	100	100	100	98.0	93.9	98.0
17	98.0	100	100	100	100	98.0	93.9	98.0
18	95.9	100	98.0	100	100	100	95.9	98.0
19	95.9	100	98.0	100	98.0	100	95.9	98.0
20	95.9	98.0	98.0	100	100	100	95.9	95.9

TABLE V
PERFORMANCE OF MUTUAL INFORMATION ON LEUKEMIA DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	94.4	94.4	94.4	94.4	94.4	94.4	94.4	94.4
2	95.8	95.8	95.8	95.8	95.8	95.8	95.8	95.8
3	95.8	95.8	95.8	95.8	95.8	95.8	97.2	97.2
4	95.8	94.4	95.8	93.1	94.4	94.4	95.8	95.8
5	95.8	95.8	94.4	94.4	95.8	95.8	95.8	95.8
6	95.8	97.2	94.4	95.8	97.2	97.2	95.8	95.8
7	95.8	95.8	94.4	95.8	94.4	94.4	95.8	93.1
8	95.8	95.8	95.8	95.8	95.8	94.4	94.4	94.4
9	95.8	95.8	97.2	98.6	94.4	95.8	95.8	95.8
10	97.2	95.8	95.8	98.6	95.8	95.8	95.8	94.4
11	94.4	95.8	95.8	97.2	94.4	95.8	95.8	93.1
12	94.4	95.8	94.4	95.8	94.4	94.4	94.4	94.4
13	95.8	95.8	93.1	95.8	93.1	93.1	95.8	94.4
14	98.6	95.8	95.8	95.8	95.8	95.8	94.4	94.4
15	97.2	95.8	93.1	93.1	93.1	93.1	94.4	93.1
16	97.2	95.8	93.1	94.4	98.6	97.2	94.4	94.4
17	97.2	95.8	94.4	97.2	97.2	95.8	95.8	95.8
18	94.4	95.8	93.1	94.4	94.4	94.4	95.8	95.8
19	94.4	95.8	93.1	94.4	94.4	94.4	95.8	94.4
20	93.1	95.8	93.1	98.6	94.4	95.8	95.8	94.4

TABLE VI
PERFORMANCE OF V-INFORMATION ON LEUKEMIA DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	94.4	90.3	90.3	94.4	94.4	94.4	94.4	94.4
2	95.8	95.8	95.8	98.6	94.4	94.4	94.4	94.4
3	97.2	100	98.6	97.2	95.8	94.4	94.4	94.4
4	95.8	98.6	97.2	97.2	95.8	95.8	95.8	95.8
5	95.8	98.6	97.2	97.2	97.2	95.8	97.2	95.8
6	95.8	100	95.8	97.2	94.4	95.8	97.2	95.8
7	95.8	98.6	97.2	97.2	94.4	97.2	97.2	94.4
8	95.8	98.6	94.4	98.6	95.8	95.8	98.6	95.8
9	98.6	98.6	93.1	98.6	95.8	95.8	98.6	94.4
10	100	98.6	97.2	98.6	97.2	95.8	98.6	97.2
11	97.2	97.2	97.2	98.6	97.2	94.4	98.6	95.8
12	98.6	98.6	97.2	98.6	95.8	94.4	98.6	95.8
13	98.6	98.6	97.2	98.6	94.4	93.1	98.6	97.2
14	98.6	98.6	95.8	97.2	94.4	94.4	95.8	95.8
15	98.6	98.6	97.2	97.2	94.4	94.4	95.8	95.8
16	100	98.6	97.2	95.8	94.4	93.1	97.2	97.2
17	97.2	98.6	95.8	95.8	94.4	95.8	95.8	95.8
18	98.6	97.2	94.4	95.8	94.4	94.4	95.8	95.8
19	100	98.6	95.8	95.8	93.1	94.4	97.2	95.8
20	98.6	98.6	95.8	95.8	91.7	94.4	97.2	94.4

TABLE VII
PERFORMANCE OF χ^2-INFORMATION ON LEUKEMIA DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	94.4	94.4	94.4	94.4	94.4	94.4	94.4	94.4
2	95.8	95.8	95.8	95.8	95.8	94.4	94.4	94.4
3	95.8	97.2	95.8	95.8	95.8	93.1	93.1	93.1
4	94.4	97.2	93.1	93.1	93.1	93.1	93.1	93.1
5	94.4	97.2	93.1	94.4	94.4	94.4	94.4	91.7
6	95.8	97.2	93.1	94.4	94.4	94.4	94.4	94.4
7	95.8	95.8	93.1	97.2	97.2	97.2	94.4	94.4
8	94.4	97.2	95.8	97.2	95.8	97.2	97.2	93.1
9	94.4	97.2	97.2	97.2	97.2	97.2	97.2	97.2
10	94.4	95.8	97.2	95.8	95.8	97.2	95.8	95.8
11	95.8	95.8	97.2	95.8	95.8	95.8	95.8	95.8
12	95.8	95.8	95.8	95.8	95.8	95.8	95.8	94.4
13	98.6	98.6	97.2	97.2	97.2	95.8	94.4	94.4
14	98.6	97.2	97.2	97.2	95.8	95.8	97.2	97.2
15	97.2	98.6	98.6	98.6	95.8	95.8	97.2	97.2
16	95.8	97.2	98.6	98.6	97.2	94.4	97.2	95.8
17	95.8	97.2	98.6	98.6	97.2	95.8	97.2	97.2
18	95.8	95.8	98.6	98.6	97.2	95.8	95.8	97.2
19	95.8	97.2	98.6	98.6	97.2	95.8	94.4	97.2
20	95.8	97.2	97.2	97.2	97.2	97.2	94.4	95.8

TABLE VIII
PERFORMANCE OF MUTUAL INFORMATION ON COLON CANCER DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	75.8	75.8	85.5	85.5	85.5	85.5	85.5	85.5
2	83.9	83.9	83.9	83.9	83.9	83.9	83.9	83.9
3	85.5	85.5	85.5	85.5	85.5	85.5	85.5	85.5
4	87.1	87.1	87.1	87.1	87.1	88.7	87.1	87.1
5	85.5	85.5	87.1	87.1	87.1	88.7	87.1	87.1
6	87.1	85.5	87.1	87.1	87.1	87.1	87.1	87.1
7	88.7	82.3	85.5	90.3	90.3	90.3	90.3	85.5
8	87.1	82.3	88.7	90.3	90.3	90.3	90.3	83.9
9	85.5	88.7	87.1	90.3	90.3	88.7	88.7	83.9
10	85.5	85.5	85.5	90.3	88.7	88.7	88.7	83.9
11	85.5	85.5	85.5	90.3	90.3	87.1	87.1	87.1
12	87.1	85.5	87.1	87.1	87.1	87.1	85.5	85.5
13	85.5	85.5	85.5	85.5	82.3	85.5	85.5	85.5
14	82.3	80.6	80.6	83.9	82.3	80.6	83.9	83.9
15	79.0	85.5	80.6	80.6	80.6	80.6	82.3	83.9
16	79.0	83.9	82.3	82.3	80.6	83.9	79.0	79.0
17	77.4	83.9	82.3	83.9	80.6	83.9	80.6	82.3
18	79.0	80.6	82.3	80.6	82.3	85.5	83.9	80.6
19	80.6	80.6	82.3	82.3	80.6	80.6	80.6	80.6
20	85.5	87.1	82.3	80.6	79.0	80.6	79.0	79.0

TABLE IX
PERFORMANCE OF V-INFORMATION ON COLON CANCER DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	85.5	85.5	85.5	85.5	85.5	85.5	85.5	85.5
2	85.5	83.9	87.1	85.5	85.5	85.5	83.9	83.9
3	87.1	87.1	85.5	85.5	88.7	85.5	88.7	85.5
4	87.1	85.5	85.5	83.9	85.5	85.5	85.5	88.7
5	88.7	90.3	85.5	90.3	87.1	87.1	88.7	88.7
6	90.3	90.3	87.1	85.5	87.1	87.1	88.7	88.7
7	88.7	93.5	87.1	82.3	85.5	85.5	90.3	90.3
8	88.7	88.7	87.1	85.5	83.9	87.1	90.3	90.3
9	85.5	90.3	87.1	87.1	88.7	85.5	90.3	90.3
10	87.1	91.9	82.3	87.1	87.1	85.5	88.7	88.7
11	82.3	90.3	88.7	87.1	83.9	85.5	88.7	88.7
12	85.5	90.3	87.1	85.5	82.3	83.9	85.5	88.7
13	88.7	88.7	88.7	80.6	80.6	83.9	85.5	88.7
14	87.1	87.1	83.9	88.7	80.6	83.9	83.9	88.7
15	85.5	87.1	83.9	88.7	80.6	83.9	83.9	85.5
16	83.9	83.9	83.9	91.9	80.6	80.6	83.9	83.9
17	83.9	80.6	83.9	90.3	80.6	80.6	83.9	80.6
18	83.9	82.3	83.9	90.3	80.6	80.6	85.5	80.6
19	83.9	82.3	83.9	91.9	79.0	80.6	83.9	83.9
20	83.9	83.9	83.9	90.3	83.9	82.3	82.3	83.9

with $\eta = 1.0$ and 13 genes with $\eta = 0.2$ and 0.5, respectively. On the other hand, V-information provides 100% accuracy using only three genes with $\eta = 0.5$. The results corresponding to leukemia data are reported in Tables V–VII. The results reported in Tables VIII–X are based on the predictive accuracy of the SVM on colon cancer data. While both V- and χ^2-information attain a maximum of 91.9% accuracy with ten genes using $\eta = 0.5$ and 20 genes using $\eta = 1.0$, respectively, mutual information provides a maximum of 90.3% accuracy using seven genes for $\eta = 1.0$ to 1.8.

Tables XI–XVI present the results of two RA data sets, namely, RAOA and RAHC. For the RAOA data set, mutual information, V-information, and χ^2-information attain 100% accuracy using three genes with $\eta = 0.2$, 0.5, and 1.2, two genes with $\eta = 1.5$, and three genes with $\eta = 0.2$ and 2.0, respectively. Similarly, 100% accuracy is obtained for the RAHC data set in the case of these three measures using seven genes with $\eta = 0.5$, six genes with $\eta = 0.2$, and 16 genes with $\eta = 1.2$, respectively. All the results reported in Tables II–XVI establish the fact that the proposed method consistently

TABLE X
PERFORMANCE OF χ^2-INFORMATION ON COLON CANCER DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	85.5	85.5	85.5	85.5	85.5	85.5	85.5	85.5
2	83.9	83.9	83.9	83.9	83.9	83.9	83.9	83.9
3	85.5	85.5	88.7	88.7	88.7	85.5	85.5	85.5
4	82.3	87.1	87.1	87.1	85.5	88.7	88.7	88.7
5	88.7	87.1	87.1	85.5	87.1	88.7	88.7	88.7
6	87.1	88.7	90.3	85.5	85.5	88.7	87.1	87.1
7	88.7	88.7	90.3	83.9	83.9	90.3	90.3	88.7
8	85.5	88.7	90.3	87.1	87.1	90.3	90.3	88.7
9	85.5	87.1	90.3	87.1	87.1	90.3	88.7	85.5
10	80.6	87.1	88.7	83.9	83.9	88.7	88.7	88.7
11	80.6	87.1	88.7	88.7	88.7	88.7	83.9	85.5
12	85.5	87.1	87.1	87.1	87.1	82.3	83.9	83.9
13	85.5	85.5	85.5	85.5	82.3	82.3	82.3	83.9
14	83.9	80.6	80.6	80.6	80.6	82.3	83.9	82.3
15	87.1	80.6	80.6	82.3	80.6	80.6	83.9	82.3
16	88.7	83.9	80.6	80.6	80.6	80.6	82.3	83.9
17	87.1	79.0	83.9	83.9	80.6	80.6	82.3	82.3
18	83.9	82.3	83.9	83.9	80.6	80.6	80.6	82.3
19	83.9	80.6	83.9	83.9	80.6	80.6	80.6	80.6
20	85.5	80.6	80.6	91.9	80.6	85.5	80.6	80.6

TABLE XI
PERFORMANCE OF MUTUAL INFORMATION ON RAOA DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	73.3	93.3	93.3	86.7	93.3	93.3	86.7	86.7
2	86.7	96.7	93.3	93.3	90.0	96.7	93.3	93.3
3	100	100	93.3	90.0	100	93.3	93.3	96.7
4	96.7	96.7	100	90.0	96.7	90.0	93.3	96.7
5	96.7	100	100	96.7	100	93.3	90.0	100
6	100	100	100	100	93.3	90.0	96.7	100
7	96.7	100	100	90.0	90.0	100	96.7	100
8	100	100	100	96.7	96.7	100	100	100
9	96.7	100	96.7	100	100	100	100	96.7
10	100	100	100	100	100	100	96.7	90.0
11	100	100	100	100	100	100	96.7	86.7
12	100	100	100	100	100	100	96.7	100
13	100	100	100	100	100	100	100	100
14	100	100	100	100	100	100	100	100
15	100	100	100	100	100	100	100	100
16	100	100	100	100	100	100	100	100
17	100	100	100	100	100	100	100	96.7
18	100	100	100	100	100	100	100	100
19	100	100	100	100	100	100	100	100
20	100	100	100	100	100	100	96.7	96.7

TABLE XII
PERFORMANCE OF V-INFORMATION ON RAOA DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	73.3	93.3	93.3	93.3	93.3	93.3	93.3	76.7
2	90.0	93.3	90.0	90.0	96.7	100	93.3	90.0
3	90.0	90.0	93.3	93.3	100	100	90.0	93.3
4	96.7	93.3	100	100	100	100	100	96.7
5	96.7	100	100	100	100	100	96.7	96.7
6	93.3	100	100	100	100	96.7	93.3	96.7
7	96.7	100	100	100	100	96.7	96.7	100
8	100	100	100	100	100	100	96.7	100
9	100	100	100	100	100	96.7	100	100
10	100	100	100	100	100	100	96.7	100
11	100	100	100	100	100	100	96.7	100
12	100	100	100	100	100	100	96.7	96.7
13	100	100	100	100	100	100	100	100
14	100	100	100	100	100	100	100	100
15	100	100	100	100	100	100	100	96.7
16	100	100	100	100	100	100	100	96.7
17	100	100	100	100	100	100	100	96.7
18	100	100	100	100	100	100	100	96.7
19	100	100	100	100	100	100	100	96.7
20	100	100	100	100	100	100	96.7	96.7

achieves better performance for $0.5 \leq \eta \leq 1.5$ with respect to the classification accuracy of the SVM, irrespective of the data sets and information measures used.

TABLE XIII
PERFORMANCE OF χ^2-INFORMATION ON RAOA DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	73.3	93.3	93.3	93.3	93.3	93.3	73.3	86.7
2	86.7	93.3	90.0	96.7	96.7	93.3	80.0	83.3
3	100	90.0	93.3	96.7	96.7	93.3	96.7	100
4	96.7	93.3	96.7	100	100	100	93.3	86.7
5	100	86.7	90.0	100	100	100	96.7	96.7
6	100	100	100	100	100	96.7	96.7	100
7	100	100	96.7	100	100	100	96.7	100
8	100	100	93.3	100	100	96.7	96.7	100
9	100	100	93.3	100	100	100	100	100
10	100	100	93.3	100	100	100	100	96.7
11	100	100	96.7	100	100	100	93.3	96.7
12	100	100	90.0	100	100	100	96.7	96.7
13	100	100	100	100	100	100	96.7	96.7
14	100	100	100	100	100	100	96.7	93.3
15	100	100	100	100	100	100	96.7	96.7
16	100	100	100	100	100	100	96.7	96.7
17	100	100	100	100	100	100	96.7	96.7
18	100	100	100	100	100	100	96.7	96.7
19	100	100	100	100	100	100	96.7	96.7
20	100	100	100	100	100	100	100	96.7

TABLE XIV
PERFORMANCE OF MUTUAL INFORMATION ON RAHC DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	78.0	78.0	78.0	78.0	78.0	78.0	78.0	78.0
2	84.0	86.0	84.0	86.0	86.0	92.0	86.0	92.0
3	82.0	92.0	90.0	90.0	82.0	92.0	94.0	94.0
4	92.0	90.0	90.0	90.0	92.0	92.0	92.0	96.0
5	90.0	88.0	94.0	94.0	92.0	88.0	92.0	92.0
6	88.0	94.0	94.0	94.0	90.0	92.0	94.0	90.0
7	90.0	100	94.0	94.0	92.0	90.0	92.0	90.0
8	98.0	96.0	96.0	96.0	96.0	90.0	90.0	86.0
9	96.0	94.0	96.0	94.0	96.0	86.0	90.0	88.0
10	92.0	94.0	94.0	92.0	92.0	92.0	90.0	90.0
11	92.0	94.0	94.0	94.0	92.0	90.0	88.0	86.0
12	90.0	90.0	96.0	94.0	94.0	90.0	92.0	94.0
13	92.0	90.0	94.0	94.0	92.0	90.0	88.0	92.0
14	94.0	92.0	100	98.0	94.0	90.0	84.0	92.0
15	92.0	92.0	96.0	96.0	96.0	86.0	86.0	90.0
16	92.0	94.0	96.0	96.0	98.0	84.0	84.0	90.0
17	92.0	92.0	96.0	98.0	98.0	90.0	90.0	88.0
18	92.0	94.0	96.0	98.0	94.0	90.0	88.0	88.0
19	94.0	92.0	98.0	98.0	98.0	90.0	86.0	88.0
20	98.0	94.0	98.0	98.0	98.0	88.0	90.0	90.0

TABLE XV
PERFORMANCE OF V-INFORMATION ON RAHC DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	78.0	84.0	84.0	78.0	78.0	80.0	80.0	76.0
2	86.0	90.0	90.0	92.0	92.0	82.0	82.0	86.0
3	90.0	90.0	96.0	92.0	92.0	86.0	88.0	92.0
4	92.0	96.0	94.0	90.0	90.0	88.0	90.0	92.0
5	98.0	96.0	94.0	84.0	84.0	92.0	90.0	92.0
6	100	96.0	94.0	88.0	96.0	98.0	96.0	94.0
7	100	96.0	94.0	92.0	92.0	98.0	96.0	96.0
8	96.0	98.0	96.0	94.0	94.0	92.0	94.0	94.0
9	94.0	100	100	100	90.0	92.0	92.0	92.0
10	94.0	96.0	100	96.0	96.0	92.0	96.0	94.0
11	98.0	98.0	100	96.0	96.0	96.0	92.0	92.0
12	98.0	94.0	100	100	94.0	96.0	92.0	94.0
13	98.0	94.0	100	100	98.0	96.0	92.0	94.0
14	100	96.0	100	100	98.0	94.0	88.0	94.0
15	100	98.0	100	100	98.0	94.0	90.0	94.0
16	100	98.0	100	100	100	92.0	88.0	96.0
17	100	100	100	100	100	94.0	90.0	96.0
18	100	100	100	100	100	94.0	90.0	96.0
19	100	100	100	100	100	92.0	90.0	96.0
20	100	100	100	100	100	96.0	96.0	100

From both the class separability and the classification accuracy analysis reported in Figs. 1–5 and Tables I–XVI, it can be seen that very large or very small amounts of overlapping among the three fuzzy sets of the input gene are found to be undesirable.

Fig. 6 represents an example scatter plot of the samples of two classes for the RAOA data set considering two top-ranked

TABLE XVI
PERFORMANCE OF χ^2-INFORMATION ON RAHC DATA

Gene/η	0.2	0.5	0.8	1.0	1.2	1.5	1.8	2.0
1	78.0	78.0	78.0	78.0	78.0	78.0	78.0	78.0
2	84.0	86.0	86.0	86.0	86.0	86.0	86.0	86.0
3	82.0	92.0	92.0	90.0	86.0	92.0	92.0	92.0
4	86.0	90.0	90.0	90.0	92.0	90.0	90.0	92.0
5	88.0	88.0	88.0	90.0	90.0	94.0	94.0	94.0
6	84.0	92.0	94.0	88.0	88.0	88.0	94.0	96.0
7	90.0	96.0	94.0	94.0	86.0	90.0	94.0	96.0
8	90.0	92.0	92.0	92.0	84.0	92.0	92.0	94.0
9	92.0	92.0	92.0	94.0	96.0	92.0	90.0	92.0
10	92.0	94.0	94.0	94.0	92.0	90.0	92.0	92.0
11	92.0	90.0	90.0	92.0	98.0	86.0	92.0	90.0
12	90.0	88.0	98.0	96.0	98.0	88.0	90.0	90.0
13	90.0	88.0	98.0	96.0	96.0	84.0	88.0	92.0
14	90.0	90.0	98.0	98.0	98.0	84.0	92.0	88.0
15	94.0	94.0	98.0	98.0	96.0	82.0	92.0	84.0
16	92.0	94.0	98.0	98.0	100	82.0	92.0	84.0
17	88.0	92.0	98.0	98.0	98.0	84.0	92.0	86.0
18	88.0	94.0	98.0	98.0	98.0	88.0	90.0	88.0
19	90.0	94.0	98.0	100	98.0	88.0	88.0	88.0
20	90.0	96.0	100	98.0	98.0	88.0	86.0	86.0

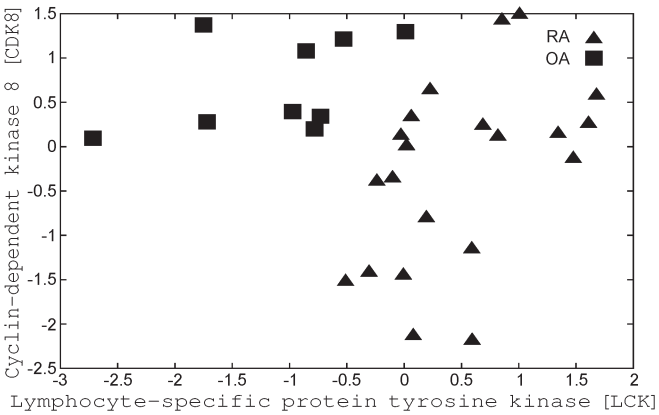


Fig. 6. Scatter plot of the samples of two classes for the RAOA data.

genes selected by the V-information measure using the proposed method at $\eta = 1.5$. From the figure, it can be seen that the samples of two classes are linearly separable.

D. Comparative Performance Analysis

Finally, Tables XVII and XVIII provide the comparative results of the different methods with respect to the classification accuracy of the SVM and the class separability index. Results are reported for five microarray data sets and three widely used information measures. From the results reported in Tables XVII and XVIII, it can be seen that the proposed method provides better or comparable classification accuracy than that of two existing methods, with a lower number of selected genes in most of the cases. However, in the case of breast cancer data using mutual information and leukemia, colon cancer, and RAHC data sets using χ^2-information, the proposed method attains the same accuracy as that of the discrete method, with a slightly higher number of genes. However, the class separability index of the 30 top-ranked genes selected by the proposed method is lower than that of the existing two methods, irrespective of the data sets and information measures used. The better performance of the proposed method is achieved due to the fact that the FEPM provides a more efficient way to approximate the true marginal and joint distributions of continuous gene expression values than the discrete and Parzen-window-based methods.

TABLE XVII
COMPARATIVE PERFORMANCE ANALYSIS ON CANCER DATA

Data Set	Measure	Method	Accuracy	Genes	$\mathcal{S}$
Breast	I	FEPM	100	8	1.889
		Discrete	100	6	2.268
		Parzen	95.9	5	2.181
	V	FEPM	100	5	2.079
		Discrete	91.8	10	2.976
		Parzen	98.0	17	3.010
	χ^2	FEPM	100	7	1.917
		Discrete	100	10	2.306
		Parzen	100	11	2.118
Leukemia	I	FEPM	98.6	9	1.374
		Discrete	98.6	19	1.604
		Parzen	98.6	12	1.613
	V	FEPM	100	3	1.357
		Discrete	100	16	1.752
		Parzen	100	7	1.686
	χ^2	FEPM	98.6	13	1.341
		Discrete	98.6	12	1.536
		Parzen	97.2	5	1.407
Colon	I	FEPM	90.3	7	3.306
		Discrete	88.7	10	4.760
		Parzen	90.3	16	4.821
	V	FEPM	91.9	10	3.334
		Discrete	91.9	12	4.850
		Parzen	90.3	8	3.985
	χ^2	FEPM	91.9	20	3.322
		Discrete	91.9	16	4.526
		Parzen	88.7	11	3.527

TABLE XVIII
COMPARATIVE PERFORMANCE ANALYSIS ON ARTHRITIS DATA

Data Set	Measure	Method	Accuracy	Genes	$\mathcal{S}$
RAOA	I	FEPM	100	3	1.767
		Discrete	100	4	2.774
		Parzen	100	3	1.992
	V	FEPM	100	2	1.683
		Discrete	100	3	3.628
		Parzen	100	3	3.704
	χ^2	FEPM	100	3	1.616
		Discrete	100	8	2.718
		Parzen	100	11	3.008
RAHC	I	FEPM	100	7	2.641
		Discrete	100	29	4.169
		Parzen	100	22	4.137
	V	FEPM	100	6	2.687
		Discrete	98.0	15	6.079
		Parzen	100	13	4.859
	χ^2	FEPM	100	16	2.576
		Discrete	100	8	3.643
		Parzen	100	23	3.892

VI. CONCLUSION AND FUTURE DIRECTION

The main contribution of this paper is threefold:

1) the development of a new concept of the FEPM to efficiently approximate the true marginal and joint distributions of continuous features;
2) the application of the proposed method in identifying discriminative and nonredundant genes from high-dimensional microarray gene expression data using different measures from information theory;
3) the comparison of the performance of the proposed method with that of two existing methods using the predictive accuracy of the SVM and the class separability index.

For three cancer and two arthritis microarray data sets, significantly better results have been found for the proposed method compared to existing methods, irrespective of the information measures used. All the results reported in this paper have demonstrated the feasibility and effectiveness of the proposed method. It is capable of identifying discriminative genes that may contribute to revealing the underlying class structures,

providing a useful tool for the exploratory analysis of biological data. The results obtained on gene microarray data sets have established that the proposed method can bring a remarkable improvement on the approximation of the true marginal and joint distributions of continuous feature values. The proposed method has only been used for the selection of genes from microarray data sets. In the future, this method will be extended to other density approximation tasks, and furthermore, its merits and limitations will be evaluated.

ACKNOWLEDGMENT

One of the authors, S. K. Pal, is a J. C. Bose Fellow of the Government of India.

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] 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.

[3] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

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

[5] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1667–1671, Dec. 2002.

[6] Y. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 52, no. 3, pp. 2318–2321, Sep. 1995.

[7] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Phys. Rev. A, Gen. Phys.*, vol. 33, no. 2, pp. 1134–1140, Feb. 1986.

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

[9] 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.

[10] P. Maji and S. K. Pal, "Rough-fuzzy C-medoids algorithm and selection of bio-basis for amino acid sequence analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 6, pp. 859–872, Jun. 2007.

[11] Q. Hu, D. Yu, Z. Xie, and J. Liu, "Fuzzy probabilistic approximation spaces and their information measures," *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 2, pp. 191–201, Apr. 2006.

[12] 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.

[13] 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.

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

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

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

[17] 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. Acad. Sci. U.S.A.*, vol. 98, no. 20, pp. 11 462–11 467, Sep. 2001.

[18] 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. Acad. Sci. U.S.A.*, vol. 96, no. 12, pp. 6745–6750, Jun. 1999.

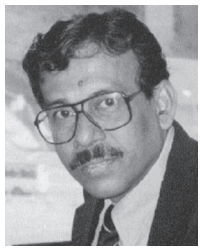
[19] T. C. T. M. van der Pouw Kraan, F. A. van Gaalen, P. V. Kasperkovitz, N. L. Verbeet, T. J. M. Smeets, M. C. Kraan, M. Fero, P.-P. Tak, T. W. J. Huizinga, E. Pieterman, F. C. Breedveld, A. A. Alizadeh, and C. L. Verweij, "Rheumatoid arthritis is a heterogeneous disease: Evidence for differences in the activation of the STAT-1 pathway between rheumatoid tissues," *Arthritis Rheum.*, vol. 48, no. 8, pp. 2132–2145, Aug. 2003.

[20] 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 patients," *Ann. Rheum. Dis.*, vol. 66, no. 8, pp. 1008–1014, Aug. 2007.

Pradipta Maji received the B.Sc.(Hons.) 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.

Currently, he is an Assistant Professor with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata. He is the author of 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, soft computing, and so forth.

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



Sankar K. Pal (M'81–SM'84–F'93) received the Ph.D. degree in radio physics and electronics from the University of Calcutta, Kolkata, India, in 1979 and the Ph.D. degree in electrical engineering, along with the Diploma of the Imperial College**,** from Imperial College, University of London, London, U.K., in 1982.

From 1986 to 1987, he was with the University of California, Berkeley, and the University of Maryland, College Park. From 1990 to 92 and in 1994, he was with the NASA Johnson Space Center, Houston, TX. In 2004, he was with the U.S. Naval Research Laboratory, Washington, DC. He held several visiting positions in Hong Kong and Australian universities. He is currently the Director and a Distinguished Scientist with the Indian Statistical Institute, Kolkata, where he founded the Machine Intelligence Unit and the Center for Soft Computing Research: A National Facility. He is a coauthor of 14 books and more than 300 research publications in the areas of pattern recognition and machine learning, image processing, data mining and web intelligence, soft computing, and bioinformatics. He is an Associate Editor and the Editor-in-Chief of many international journals.

Dr. Pal is a Fellow of the Academy of Sciences for the Developing World (TWAS), Italy, the International Association for Pattern recognition, U.S., the International Association of Fuzzy Systems, U.S., and all the four National Academies for Science/Engineering in India. Since 1997, he has been serving as a Distinguished Visitor of the IEEE Computer Society (U.S.) for the Asia-Pacific Region. He is the recipient of the 1990 S.S. Bhatnagar Prize and many prestigious awards in India and abroad, including the 1999 G.D. Birla Award, the 1998 Om Bhasin Award, the 1993 Jawaharlal Nehru Fellowship, the 2000 Khwarizmi International Award from the Islamic Republic of Iran, the 2000–2001 FICCI Award, the 1993 Vikram Sarabhai Research Award, the 1993 NASA Tech Brief Award (U.S.), the 1994 IEEE TRANSACTIONS ON NEURAL NETWORKS Outstanding Paper Award (U.S.), the 1995 NASA Patent Application Award (U.S.), the 1997 IETE-R.L. Wadhwa Gold Medal, the 2001 INSA-S.H. Zaheer Medal, the 2005–2006 ISC-P.C. Mahalanobis Birth Centenary Award (Gold Medal) for Lifetime Achievement, the 2007 J.C. Bose Fellowship of the Government of India, and the 2008 Vigyan Ratna Award from Science and Culture Organization, West Bengal.