Fuzzy–Rough Simultaneous Attribute Selection and Feature Extraction Algorithm

Pradipta Maji and Partha Garai

Abstract—**Among the huge number of attributes or features present in real-life data sets, only a small fraction of them are effective to represent the data set accurately. Prior to analysis of the data set, selecting or extracting relevant and significant features is an important preprocessing step used for pattern recognition, data mining, and machine learning. In this regard, a novel dimensionality reduction method, based on fuzzy–rough sets, that simultaneously selects attributes and extracts features using the concept of feature significance is presented. The method is based on maximizing both the relevance and significance of the reduced feature set, whereby redundancy therein is removed. This paper also presents classical and neighborhood rough sets for computing the relevance and significance of the feature set and compares their performances with that of fuzzy–rough sets based on the predictive accuracy of nearest neighbor rule, support vector machine, and decision tree. An important finding is that the proposed dimensionality reduction method based on fuzzy–rough sets is shown to be more effective for generating a relevant and significant feature subset. The effectiveness of the proposed fuzzy–rough-set-based dimensionality reduction method, along with a comparison with existing attribute selection and feature extraction methods, is demonstrated on real-life data sets.**

Index Terms—**Attribute selection, classification, feature extraction, pattern recognition, rough sets.**

I. Introduction

DIMENSIONALITY reduction is a process of selecting a map by which a sample in an m-dimensional measurement space is transformed into an object in a d-dimensional feature space, where $d < m$ [1]. The main objectives of this task are to retain or generate the optimum salient characteristics necessary for the pattern recognition process and to reduce the dimensionality of the measurement space so that effective and easily computable algorithms can be devised for efficient clustering or classification [2].

The problem of dimensionality reduction has two aspects, namely, formulation of a suitable criterion to evaluate the goodness of a feature set and search of the optimal set in terms of the criterion [3]. In general, those features are considered to have optimal saliencies for which interclass (intraclass) distances are maximized (minimized). The criterion of a good feature is that it should be unchanging with any other possible variation within a class while emphasizing differences that are important in discriminating between patterns of different classes [4].

The major mathematical measures so far devised for the estimation of feature quality are mostly statistical in nature and can be broadly classified into two categories, namely, feature selection in the measurement space and feature selection in a transformed space. The techniques in the first category generally reduce the dimensionality of the measurement space by discarding redundant or least information-carrying features [5]. On the other hand, those in the second category utilize all the information contained in the measurement space to obtain a new transformed space, thereby mapping a higher dimensional pattern to a lower dimensional one. This is referred to as feature extraction [1], [2].

An optimal feature subset selected or extracted by a dimensionality reduction method is always relative to a certain feature evaluation criterion. In general, different criteria may lead to different optimal feature subsets. However, every criterion tries to measure the discriminating ability of a feature or a subset of features to distinguish different class labels. One of the main problems in real-life 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 rough set theory [6] has gained popularity in modeling and propagating uncertainty. It has been applied to reasoning with uncertainty, fuzzy rule extraction and modeling, classification, clustering, and feature selection [6], [7].

Rough sets can be used to find the most informative feature subset of original attributes from a given data set with discretized attribute values [8]–[10]. However, there are usually real-valued data and fuzzy information in real-world applications. In rough sets, the real-valued features are divided into several discrete partitions, and the dependence or the quality of approximation of a feature is calculated. The inherent error that exists in the discretization process is of major concern in the computation of the dependence of real-valued features. Combining fuzzy and rough sets provides an important direction in reasoning with uncertainty for real-valued data [11]–[13]. They are complementary in some aspects. The generalized theories of rough–fuzzy computing have been applied successfully to feature selection of real-valued data sets [7], [11], [14]–[18]. Also, neighborhood rough sets [19], [20] are found to be suitable for both numerical and categorical data sets. In [19], Hu *et al.* described a neighborhood-rough-set-based feature selection algorithm.

On the other hand, a feature extraction technique such as principal component analysis (PCA), linear discriminant analysis, and independent component analysis [2] generates a new

The authors are with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India (e-mail: pmaji@isical.ac.in; parthagarai_r@isical.ac.in).

set of features using a mapping function that takes some linear or nonlinear combination of original features. While PCA uses a linear orthogonal transformation to project a sample space containing possibly correlated variables into a different space with uncorrelated variables, independent component analysis decomposes a multidimensional feature vector into statistically independent components to reveal the hidden factors from a set of random variables [2].

In general, a feature extraction technique provides a feature subset richer than that obtained using a feature selection algorithm with a higher cost [21]. Hence, it is very difficult to decide whether to select a feature from the original measurement space or to extract a new feature by transforming the existing features for a given data set. A dimensionality reduction algorithm that can simultaneously select and extract features depending upon the criteria needs to be formulated, integrating the merits of both feature selection and extraction techniques.

In this regard, a novel dimensionality reduction algorithm is proposed based on fuzzy–rough sets, which simultaneously selects and extracts features from a given data set. Using the concept of feature significance, the feature set in each iteration is partitioned into three subsets, namely, insignificant, dispensable, and significant feature sets. The insignificant feature set is discarded from the current feature set, while the significant feature set is used to select or extract a feature in the next iteration. Depending on the quality of features present in the dispensable set of the current iteration, a new feature is extracted or an existing feature is selected from the dispensable set for a reduced feature set. In effect, the final reduced feature set may simultaneously contain some original features of the measurement space and extracted new features of the transformed space, which are both relevant and significant. The effectiveness of the proposed fuzzy–rough dimensionality reduction method, along with a comparison with other methods, is demonstrated on a set of real-life data using the predictive accuracy of nearest neighbor rule, support vector machine (SVM), and decision tree.

The structure of the rest of this paper is as follows. Section II briefly introduces the basic notions of rough sets, neighborhood rough sets, and fuzzy–rough sets. The proposed fuzzy–rough-set-based simultaneous attribute selection and feature extraction method is described in Section III. A few case studies and a comparison with other methods are presented in Section IV. Concluding remarks are given in Section V.

II. Different Rough Set Models

In this section, the basic notions in the theories of rough sets, neighborhood 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 $\mathbb{U} = \{x_1, \ldots, x_i, \ldots, x_n\}$ is a nonempty set, also called 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 $f : \mathbb{U} \times \mathbb{A} \to V$. An approximation space is also called an information system [6].

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

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

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

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

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 $\mathrm{IND}(\mathbb{P})$ 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 [6]:

$$\begin{aligned} \underline{\mathbb{P}}(X) &= \bigcup \{[x_i]_{\mathbb{P}} | [x_i]_{\mathbb{P}} \subseteq X\} \\ \overline{\mathbb{P}}(X) &= \bigcup \{[x_i]_{\mathbb{P}} | [x_i]_{\mathbb{P}} \cap X \neq \emptyset\}. \end{aligned} \tag{2}$$

Hence, the lower approximation $\underline{\mathbb{P}}(X)$ is the union of all elementary sets which are subsets of X, and the upper approximation $\overline{\mathbb{P}}(X)$ is the union of all elementary sets which 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 set of X. The lower (upper) approximation $\underline{\mathbb{P}}(X)$ $[\overline{\mathbb{P}}(X)]$ is interpreted as the collection of those elements of $\mathbb{U}$ that definitely (possibly) belong to X. The lower approximation is also called positive region sometimes, denoted as $\mathrm{POS}_{\mathbb{P}}(X)$. A set X is said to be definable 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 $\mathbb{C}$ is the condition attribute set and $\mathbb{D}$ is the decision attribute set. The dependence between $\mathbb{C}$ and $\mathbb{D}$ can be defined as

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|\mathrm{POS}_{\mathbb{C}}(\mathbb{D})|}{|\mathbb{U}|} \tag{3}$$

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

B. Neighborhood Rough Sets

Given an arbitrary object $x_i \in \mathbb{U}$ and $\mathbb{P} \subseteq \mathbb{C}$, the neighborhood $\Phi_{\mathbb{P}}(x_i)$ of x_i with given threshold Φ, in feature space $\mathbb{P}$, is defined as [19]

$$\Phi_{\mathbb{P}}(x_i) = \left\{x_j | x_j \in \mathbb{U}, \Delta^{\mathbb{P}}(x_i, x_j) \leq \Phi\right\} \tag{4}$$

where Δ is a distance function. $\Phi_{\mathbb{P}}(x_i)$ in (4) is the neighborhood information granule centered with sample x_i. The neighborhood granule generation is effected by two key factors,

namely, the used distance function Δ and parameter Φ. The first one determines the shape and the second controls the size of the neighborhood granule. Both these factors play important roles in neighborhood rough sets and can be considered to control the granularity of data analysis. The significance of attributes varies with the granularity levels. Accordingly, the neighborhood-rough-set-based algorithm selects different attribute subsets with the change of the Δ function and the Φ value [19], [20].

Hence, each sample generates granules with a neighborhood relation. For a metric space $\langle \mathbb{U}, \Delta \rangle$, the set of neighborhood granules $\{\Phi(x_i)|x_i \in \mathbb{U}\}$ forms an elemental granule system that covers the universal space rather than partitions it as in the case of rough sets. It is noted that the partition of space generated by rough sets can be obtained from neighborhood rough sets with covering principle, while the other way around is not possible. Moreover, a neighborhood granule degrades to an equivalence class for $\Phi = 0$. In this case, the samples in the same neighborhood granule are equivalent to each other, and the neighborhood rough set model degenerates to rough sets. Hence, neighborhood rough sets can be treated as a generalized case of rough sets.

C. 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 or fuzzy knowledge granules. This means that the decision and condition attributes may all be fuzzy [12].

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 upper approximations are then defined as follows [12]:

$$\mu_{\underline{\mathbb{P}}X}(F_i) = \inf_x \{\max\{(1 - \mu_{F_i}(x)), \mu_X(x)\}\} \qquad \forall i \tag{5}$$

$$\mu_{\overline{\mathbb{P}}X}(F_i) = \sup_x \{\min\{\mu_{F_i}(x), \mu_X(x)\}\} \qquad \forall i \tag{6}$$

where F_i represents a fuzzy equivalence class belonging to $\mathbb{U}/\mathbb{P}$, the partition of $\mathbb{U}$ generated by $\mathbb{P}$, 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 of this, the fuzzy lower and upper approximations can be defined as [11]

$$\mu_{\underline{\mathbb{P}}X}(x) = \sup_{F_i \in \mathbb{U}/\mathbb{P}} \min\left\{\mu_{F_i}(x), \mu_{\underline{\mathbb{P}}X}(F_i)\right\} \tag{7}$$

$$\mu_{\overline{\mathbb{P}}X}(x) = \sup_{F_i \in \mathbb{U}/\mathbb{P}} \min\left\{\mu_{F_i}(x), \mu_{\overline{\mathbb{P}}X}(F_i)\right\}. \tag{8}$$

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

$$\mu_{\mathrm{POS}_{\mathbb{C}}(\mathbb{D})}(x) = \sup_{X \in \mathbb{U}/\mathbb{D}} \mu_{\underline{\mathbb{C}}X}(x) \tag{9}$$

where $\mathbb{A} = \mathbb{C} \cup \mathbb{D}$. Using the definition of fuzzy positive region, the dependence function can be defined as follows [11]:

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|\mu_{\mathrm{POS}_{\mathbb{C}}(\mathbb{D})}(x)|}{|\mathbb{U}|} = \frac{1}{|\mathbb{U}|} \sum_{x \in \mathbb{U}} \mu_{\mathrm{POS}_{\mathbb{C}}(\mathbb{D})}(x). \tag{10}$$

III. Proposed Dimensionality Reduction Method

In this section, a new dimensionality reduction method is presented, integrating the theory of fuzzy–rough sets and the merits of both feature selection and extraction techniques. It is based on the concept of feature significance that follows next.

A. Feature Significance

In real-life data analysis, one of the important issues is computing both relevance and redundancy of features by discovering dependences among them. Intuitively, a set of features $\mathbb{B}$ depends totally on a set of features $\mathbb{A}$, if all feature values from $\mathbb{B}$ are uniquely determined by the values of the features from $\mathbb{A}$. If there exists a functional dependence between the values of $\mathbb{B}$ and $\mathbb{A}$, then $\mathbb{B}$ depends totally on $\mathbb{A}$.

Let $\mathbb{U} = \{x_1, \ldots, x_i, \ldots, x_n\}$ be the set of n samples and $\mathbb{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_j, \ldots, \mathcal{A}_m\}$ denote the set of m features of a given data set. Define $\gamma_{\mathcal{A}_i}(\mathbb{D})$ as the relevance of the feature $\mathcal{A}_i$ with respect to the class label or decision attribute $\mathbb{D}$. The relevance represents the quality of a feature or the degree of dependence of decision attribute $\mathbb{D}$ on condition attribute $\mathcal{A}_i$. Any rough set model reported earlier can be used to compute the relevance of a feature or a set of features.

To what extent a feature is contributing to calculate the joint relevance or dependence can be calculated by the significance of that feature. The change in dependence when a feature is removed from the set of features is a measure of the significance of the feature.

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

$$\sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathcal{A}_j, \mathbb{D}) = \gamma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}) - \gamma_{\mathcal{A}_i}(\mathbb{D}). \tag{11}$$

Hence, the significance of a feature $\mathcal{A}_j$ is the change in dependence when the feature $\mathcal{A}_j$ is removed from the set $\{\mathcal{A}_i, \mathcal{A}_j\}$. The higher the change in dependence, the more significant the feature $\mathcal{A}_j$. If the significance is zero, then the feature $\mathcal{A}_j$ is dispensable. The following properties can be stated about the measure.

1) $\sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathcal{A}_j, \mathbb{D}) = 0$ if and only if the feature $\mathcal{A}_j$ is dispensable in the set $\{\mathcal{A}_i, \mathcal{A}_j\}$.
2) $\sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathcal{A}_j, \mathbb{D}) < 0$ if the feature $\mathcal{A}_i$ is more relevant than the feature set $\{\mathcal{A}_i, \mathcal{A}_j\}$.
3) $\sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathcal{A}_j, \mathbb{D}) > 0$ if the feature $\mathcal{A}_j$ is significant with respect to another feature $\mathcal{A}_i$.
4) $\sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathcal{A}_j, \mathbb{D}) \neq \sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathcal{A}_i, \mathbb{D})$ (asymmetric).

B. Simultaneous Feature Selection and Extraction

A high-dimensional real-life data set generally may contain a number of nonrelevant and insignificant features. The presence of such features may lead to a reduction in the useful information. Ideally, the reduced feature set obtained using

a dimensionality reduction algorithm should contain features that have high relevance with the classes while the significance among them would be as high as possible. The relevant and significant features are expected to be able to predict the classes of the samples. Hence, to assess the effectiveness of the features, both relevance and significance need to be measured quantitatively. The proposed dimensionality reduction method addresses the aforementioned issues through the following three phases:

1) computation of the relevance of each feature present in the original feature set;
2) determination of the insignificant, dispensable, and significant feature sets;
3) extraction of a relevant feature from the dispensable set.

The fuzzy–rough set is used to compute both the relevance and significance of features. The insignificant feature set is discarded from the whole feature set, while the significant feature set is used to select or extract significant features for a reduced feature set.

Let $\gamma_{\mathcal{A}_i}(\mathbb{D})$ represent the relevance of feature $\mathcal{A}_i \in \mathbb{C}$. The proposed algorithm starts with a single feature $\mathcal{A}_i$ that has the highest relevance value. Based on the significance values of all other features, the feature set $\mathbb{C}$ is then partitioned into three subsets, namely, insignificant set I_i, dispensable set D_i, and significant set S_i, which are defined as follows:

$$I_i = \left\{\mathcal{A}_j | \sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathcal{A}_j, \mathbb{D}) < -\delta_i; \mathcal{A}_j \neq \mathcal{A}_i \in \mathbb{C}\right\} \quad (12)$$

$$D_i = \left\{\mathcal{A}_j | -\delta_i \leq \sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathcal{A}_j, \mathbb{D}) \leq \delta_i; \mathcal{A}_j \neq \mathcal{A}_i \in \mathbb{C}\right\} \quad (13)$$

$$S_i = \left\{\mathcal{A}_j | \sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathcal{A}_j, \mathbb{D}) > \delta_i; \mathcal{A}_j \neq \mathcal{A}_i \in \mathbb{C}\right\} \quad (14)$$

where $\mathbb{C} = I_i \cup D_i \cup S_i \cup \{\mathcal{A}_i\}$ and δ_i is a predefined threshold value corresponding to the feature $\mathcal{A}_i$.

The insignificant set I_i represents the set of features that are insignificant with respect to the candidate feature $\mathcal{A}_i$ of the current iteration. Hence, the insignificant set I_i should be discarded from the whole feature set $\mathbb{C}$ as the presence of such insignificant features may lead to a reduction in the useful information. If insignificant features are present in the reduced feature set, they may reduce the classification or clustering performance. The significant set S_i consists of a set of features that are significant with respect to the feature $\mathcal{A}_i$. In other words, the set S_i represents the set of features of $\mathbb{C}$ that have the significance values with respect to the feature $\mathcal{A}_i$ greater than the threshold δ_i. This set is considered in the next iteration to select or extract a new feature.

On the other hand, the dispensable set D_i is used for extracting a new feature in the current iteration. As the significance values of the features present in the dispensable set are very low, they form a group of similar features. These features may be considered to generate a new feature. However, the similar features of the dispensable set may be in phase or out of phase with respect to each other. Hence, the following definition can be used to extract a new feature $\overline{\mathcal{A}}_i$ from the dispensable set of features D_i:

$$\overline{\mathcal{A}}_i = \frac{\mathcal{A}_i + \sum \lambda_j \mathcal{A}_j}{1 + \sum |\lambda_j|} \quad (15)$$

where $\lambda_j \in \{-1, 0, 1\}$ and $\mathcal{A}_j \in D_i$.

To find out the value of λ_j for each feature $\mathcal{A}_j \in D_i$, the following greedy algorithm can be used. Let $\mathcal{A}_i$ be the initial representative of the set D_i. The representative of D_i is refined incrementally. By searching among the features of set D_i, the current representative is merged and averaged with other features, both in phase and out of phase, such that the augmented representative $\overline{\mathcal{A}}_i$ increases the relevance value. The merging process is repeated until the relevance value can no longer be improved. If a feature $\mathcal{A}_j \in D_i$ in phase (out of phase) with the feature $\mathcal{A}_i$ increases the relevance value, then $\lambda_j = 1$ ($\lambda_j = -1$). On the other hand, the value of $\lambda_j = 0$ if feature $\mathcal{A}_j$ does not increase the relevance value, irrespective of the phases. The main steps to find out the values of λ_j for all $\mathcal{A}_j \in D_i$ are as follows.

1) Initialize $\bar{D}_i \leftarrow \{\mathcal{A}_i\}$, $\gamma_{\max} \leftarrow \gamma_{\mathcal{A}_i}$, and $\lambda_i = 1$.
2) Repeat the following six steps (steps 3–8) until $D_i = \emptyset$.
3) Initialize $\max \leftarrow 0$ and $\lambda_j = 0\ \forall \mathcal{A}_j \in D_i$.
4) Repeat the following three steps (steps 5–7) for each feature $\mathcal{A}_j \in D_i$.
5) Compute two augmented representatives $\mathcal{A}^+_{i+j}$ and $\mathcal{A}^-_{i+j}$ by averaging the features of $\bar{D}_i$ with $\mathcal{A}_j$ and its complement, respectively, as follows:

$$\mathcal{A}^+_{i+j} = \frac{1}{1 + \sum |\lambda_k|} \left\{ \sum_{\mathcal{A}_k \in \bar{D}_i} \lambda_k \mathcal{A}_k + \mathcal{A}_j \right\} \quad (16)$$

$$\mathcal{A}^-_{i+j} = \frac{1}{1 + \sum |\lambda_k|} \left\{ \sum_{\mathcal{A}_k \in \bar{D}_i} \lambda_k \mathcal{A}_k - \mathcal{A}_j \right\}. \quad (17)$$

6) Evaluate the value of λ_j as follows:

$$\lambda_j = \begin{cases} 1, & \text{if } \gamma_{\mathcal{A}^+_{i+j}} \geq \gamma_{\mathcal{A}^-_{i+j}} \text{ and } \gamma_{\mathcal{A}^+_{i+j}} > \gamma_{\max} \\ -1, & \text{if } \gamma_{\mathcal{A}^-_{i+j}} \geq \gamma_{\mathcal{A}^+_{i+j}} \text{ and } \gamma_{\mathcal{A}^-_{i+j}} > \gamma_{\max}. \end{cases}$$

7) If $\lambda_j \neq 0$, then $\max \leftarrow j$ and

$$\gamma_{\max} = \begin{cases} \gamma_{\mathcal{A}^+_{i+j}}, & \text{if } \lambda_j = 1 \\ \gamma_{\mathcal{A}^-_{i+j}}, & \text{if } \lambda_j = -1. \end{cases} \quad (18)$$

8) If $\max \neq 0$, then $D_i \leftarrow D_i \setminus \{\mathcal{A}_{\max}\}$ and $\bar{D}_i \leftarrow \bar{D}_i \cup \{\mathcal{A}_{\max}\}$; otherwise, stop.

After extracting the feature $\overline{\mathcal{A}}_i$ from the dispensable set D_i using (15), the insignificant feature set I_i and the used features of D_i are discarded from the whole feature set $\mathbb{C}$. From the remaining features of $\mathbb{C}$, another feature $\mathcal{A}_j$ is selected by maximizing the following condition:

$$\gamma_{\mathcal{A}_j}(\mathbb{D}) + \frac{1}{|\mathbb{S}|} \sum_{\overline{\mathcal{A}}_i \in \mathbb{S}} \sigma_{\{\overline{\mathcal{A}}_i,\mathcal{A}_j\}}(\mathcal{A}_j, \mathbb{D}) \quad (19)$$

where $\mathbb{S}$ is the already selected or extracted feature set. The process is repeated to select or extract more features. The main steps of the proposed simultaneous feature selection and extraction algorithm are reported as follows.

- *Input:* Original set $\mathbb{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_k, \ldots, \mathcal{A}_m\}$.
- *Output:* Reduced set $\mathbb{S} = \{\overline{\mathcal{A}}_1, \ldots, \overline{\mathcal{A}}_i, \ldots, \overline{\mathcal{A}}_d\}$.

1) Initialize $\mathcal{B} \leftarrow \{\mathcal{A}_1, \ldots, \mathcal{A}_k, \ldots, \mathcal{A}_m\}$ and $\mathbb{S} \leftarrow \emptyset$.
2) Calculate relevance value $\gamma_{\mathcal{A}_i}(\mathbb{D})$ of feature $\mathcal{A}_i \in \mathcal{B}$.

3) Select feature $\mathcal{A}_i$ from $\mathcal{B}$ as the first feature that has the highest relevance value.
4) Repeat the following five steps (steps a–e) until $\mathcal{B} = \emptyset$ or desired number of features are selected.
   a) Generate three subsets, namely, insignificant set I_i, dispensable set D_i, and significant set S_i, with respect to the candidate feature $\mathcal{A}_i$.
   b) Evaluate the values of λ_j for all $\mathcal{A}_j \in D_i$.
   c) Extract feature $\overline{\mathcal{A}}_i$ from the dispensable set D_i using (15), and add it to the reduced set $\mathbb{S}$.
   d) Discard the subset I_i and the used features of D_i from the candidate feature set $\mathcal{B}$.
   e) From the remaining features of $\mathcal{B}$, select feature $\mathcal{A}_j$ that maximizes condition (19).
5) Stop.

C. Fundamental Property

From the earlier discussions, the following properties can be stated about the proposed algorithm.

1) The relevance of extracted feature $\overline{\mathcal{A}}_i$ is higher than that of original feature $\mathcal{A}_i$, i.e., $\gamma_{\overline{\mathcal{A}}_i}(\mathbb{D}) \geq \gamma_{\mathcal{A}_i}(\mathbb{D})$.
2) The significance $\sigma_{\{\overline{\mathcal{A}}_i, \overline{\mathcal{A}}_j\}}(\overline{\mathcal{A}}_i, \mathbb{D})$ between any two features $\overline{\mathcal{A}}_i$ and $\overline{\mathcal{A}}_j$ of reduced feature set $\mathbb{S}$ is high.
3) If $\lambda_j = 0$ for all $\mathcal{A}_j \in D_i$, then the extracted feature $\overline{\mathcal{A}}_i$ at a particular iteration is actually the candidate feature $\mathcal{A}_i$ of the original feature set $\mathbb{C}$.

Hence, the proposed dimensionality reduction method generates a reduced feature set $\mathbb{S}$ that may simultaneously contain some selected features of the original measurement space and some extracted features of the transformed feature space, which are both relevant and significant.

D. Computational Complexity

The proposed fuzzy–rough simultaneous feature selection and extraction algorithm has low computational complexity with respect to the number of features present in the original set.

The computation of the relevance of m features is carried out in step 2 of the proposed algorithm, which has $\mathcal{O}(m)$ time complexity. The selection of the most relevant feature from the set of m features, which is carried out in step 3, has also a complexity $\mathcal{O}(m)$. There is only one loop in step 4 of the proposed dimensionality reduction method, which is executed $(d-1)$ times, where d represents the number of features in the reduced feature set. To generate insignificant, dispensable, and significant sets, the significance of all existing features needs to be computed. The computation of the significance of a feature with respect to another feature takes only a constant amount of time. If $\acute{m} < m$ represents the cardinality of the existing feature set, the complexity to compute the significance of $\acute{m}$ features, which is carried out in step 4-a, is $\mathcal{O}(\acute{m})$. The computational complexity of extracting a new feature from the dispensable feature set with cardinality $|D_i| < \acute{m}$, which is carried out in steps 4-b and 4-c, is given by $\mathcal{O}(|D_i|^2)$. The computation of the significance of $(\acute{m} - |D_i| - |I_i|)$ candidate features with respect to the already-selected features, which is carried out in step 4-e, has a complexity $\mathcal{O}(\acute{m} - |D_i| - |I_i|)$, where $|I_i|$ is the cardinality of the insignificant set. In effect, the selection of a feature from $(\acute{m} - |D_i| - |I_i|)$ candidate features by maximizing both relevance and significance has also a complexity $\mathcal{O}(\acute{m} - |D_i| - |I_i|)$.

Hence, the total complexity to execute the loop $(d-1)$ times is $\mathcal{O}((d-1)(\acute{m} + |D_i|^2 + (\acute{m} - |D_i| - |I_i|)))$. In effect, the selection of a set of d features from the whole set of m features using the proposed simultaneous feature selection and extraction algorithm has an overall computational complexity of $\mathcal{O}(m) + \mathcal{O}(m) + \mathcal{O}((d-1)(\acute{m} + |D_i|^2 + (\acute{m} - |D_i| - |I_i|))) \simeq \mathcal{O}(m + (d-1)(\acute{m} + |D_i|^2 - |I_i|))$, where $|D_i| < \acute{m} < m$ and $|I_i| < \acute{m} < m$.

E. Selection of Threshold

The threshold δ_i in (13) plays an important role to form the dispensable set corresponding to the candidate feature $\mathcal{A}_i$ at a particular iteration. It controls the degree of similarity among the features in the dispensable set. In effect, it has a direct influence on the performance of the proposed simultaneous feature selection and extraction algorithm. If δ_i increases, the number of features or attributes in the dispensable set increases, but the similarity among them with respect to sample categories or class labels decreases. The similarity among the features in the dispensable set increases with a decrease of δ_i.

To find out the optimum value of δ_i, the following definition, based on the significance values of the candidate feature set $\mathcal{B}$ for each iteration, can be used:

$$\delta_i = \left[\frac{1}{|\mathcal{B}|} \sum_{\mathcal{A}_j \in \mathcal{B}} \left\{ \sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathcal{A}_j, \mathbb{D}) \right\}^r \right]^{\frac{1}{r}} \tag{20}$$

where r is a positive integer. Hence, the threshold δ_i represents the zero-mean rth-order moment of the significance values of the attributes $\mathcal{A}_j \in \mathcal{B}$ with respect to the candidate feature $\mathcal{A}_i$.

F. Computation of Relevance and Significance

In the proposed dimensionality reduction method, both the relevance and significance of a set of features are computed using fuzzy equivalence partition matrix that follows next.

1) Fuzzy Equivalence Partition Matrix: Given a finite set $\mathbb{U}$, $\mathbb{C}$ 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 c partitions of $\mathbb{U}$ can be arrayed as a $(c \times n)$ matrix $\mathbb{M}_{\mathbb{C}}$, termed as fuzzy equivalence partition matrix [17], which is denoted by

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

subject to $\sum_{i=1}^{c} m_{ij}^{\mathbb{C}} = 1 \ \forall j$ and for any value of i; if $k = \arg\ \max_j \{m_{ij}^{\mathbb{C}}\}$, then $\max_j \{m_{ij}^{\mathbb{C}}\} = \max_l \{m_{lk}^{\mathbb{C}}\} > 0$, where

$m_{ij}^{\mathbb{C}} \in [0,1]$ represents the membership of object x_j in the ith fuzzy equivalence partition or class F_i. Using the concept of fuzzy equivalence partition matrix, the dependence between condition attribute set $\mathbb{C}$ and decision attribute set $\mathbb{D}$ can be redefined as follows [17]:

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{1}{n}\sum_{j=1}^{n}\kappa_j \tag{22}$$

$$\kappa_j = \sup_k\left\{\sup_i\left\{\min\left\{m_{ij}^{\mathbb{C}}, \inf_l\left\{\max\left\{1-m_{il}^{\mathbb{C}}, m_{kl}^{\mathbb{D}}\right\}\right\}\right\}\right\}\right\}. \tag{23}$$

2) Generation of Fuzzy Equivalence Partition: The family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes [12], [17]. In general, the π function in the 1-D form is used to assign membership values to different fuzzy equivalence classes for the input features [22], [23]. A fuzzy set with membership function $\pi(x;\bar{c},\sigma)$ [24] represents a set of points clustered around $\bar{c}$, where

$$\pi(x;\bar{c},\sigma) = \begin{cases} 2\left(1-\frac{\|x-\bar{c}\|}{\sigma}\right)^2, & \text{for } \frac{\sigma}{2} \le \|x-\bar{c}\| \le \sigma \\ 1-2\left(\frac{\|x-\bar{c}\|}{\sigma}\right)^2, & \text{for } 0 \le \|x-\bar{c}\| \le \frac{\sigma}{2} \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

where $\sigma > 0$ is the radius of the π function with $\bar{c}$ as the central point and $\|\cdot\|$ denotes the Euclidean norm. 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, i.e., $\pi(\bar{c};\bar{c},\sigma) = 1$. The membership value of a point decreases as its distance from the central point $\bar{c}$, i.e., $\|x-\bar{c}\|$, increases. When $\|x-\bar{c}\| = (\sigma/2)$, the membership value of x is 0.5, and this is called a crossover point [24].

Each real-valued feature 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 feature axis can be determined automatically from the distribution of training patterns or objects [24]. Let $\bar{m}_i$ be the mean of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$ along the ith feature $\mathcal{A}_i$. Then, $\bar{m}_{i_l}$ and $\bar{m}_{i_h}$ are defined as the means, along the ith feature, of the objects having coordinate values in the ranges of $[\mathcal{A}_{i_{\min}}, \bar{m}_i)$ 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 feature $\mathcal{A}_i$ for the training set. For three fuzzy sets low, medium, and high, the centers and corresponding radii are as follows [24]:

$$\bar{c}_{\text{low}}(\mathcal{A}_i) = \bar{m}_{i_l}\bar{c}_{\text{medium}}(\mathcal{A}_i) = \bar{m}_i\bar{c}_{\text{high}}(\mathcal{A}_i) = \bar{m}_{i_h} \tag{25}$$

$$\begin{aligned} \sigma_{\text{low}}(\mathcal{A}_i) &= 2\left(\bar{c}_{\text{medium}}(\mathcal{A}_i) - \bar{c}_{\text{low}}(\mathcal{A}_i)\right) \\ \sigma_{\text{high}}(\mathcal{A}_i) &= 2\left(\bar{c}_{\text{high}}(\mathcal{A}_i) - \bar{c}_{\text{medium}}(\mathcal{A}_i)\right) \end{aligned} \tag{26}$$

$$\begin{aligned} \sigma_{\text{medium}}(\mathcal{A}_i) = {} & \frac{\eta}{(\mathcal{A}_{i_{\max}} - \mathcal{A}_{i_{\min}})} \\ & \times \left[\sigma_{\text{low}}(\mathcal{A}_i)\left(\mathcal{A}_{i_{\max}} - c_{\text{medium}}(\mathcal{A}_i)\right)\right. \\ & \left. + \sigma_{\text{high}}(\mathcal{A}_i)\left(c_{\text{medium}}(\mathcal{A}_i) - \mathcal{A}_{i_{\min}}\right)\right] \end{aligned} \tag{27}$$

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

A $c \times n$ fuzzy equivalence partition matrix $\mathbb{M}_{\mathbb{C}}$ represents the c fuzzy equivalence partitions of the universe generated by a fuzzy equivalence relation. Each row of the matrix $\mathbb{M}_{\mathbb{C}}$ is a fuzzy equivalence partition or class. The $c \times n$ fuzzy equivalence partition matrix $\mathbb{M}_{\mathcal{A}_i}$, corresponding to the ith feature $\mathcal{A}_i$, can be calculated from the c fuzzy equivalence classes of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$, where

$$m_{kj}^{\mathcal{A}_i} = \frac{\pi(x_j;\bar{c}_k,\sigma_k)}{\sum_{l=1}^{c}\pi(x_j;\bar{c}_l,\sigma_l)}. \tag{28}$$

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

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

In effect, each position $m_{kj}^{\mathcal{A}_i}$ of the fuzzy equivalence partition matrix $\mathbb{M}_{\mathcal{A}_i}$ must satisfy the following conditions [17]: $m_{kj}^{\mathcal{A}_i} \in [0,1]$; $\sum_{k=1}^{c} m_{kj}^{\mathcal{A}_i} = 1\ \forall j$ and for any value of k, if $s = \arg\max_j\{m_{kj}^{\mathcal{A}_i}\}$, then $\max_j\{m_{kj}^{\mathcal{A}_i}\} = \max_l\{m_{ls}^{\mathcal{A}_i}\} > 0$.

IV. Experimental Results

The performance of the proposed fuzzy–rough simultaneous attribute selection and feature extraction method is extensively studied and compared with those of some existing feature selection and extraction algorithms, namely, maximal-relevance (Max-Relevance) and maximal-relevance maximal-significance (MRMS) [9] frameworks with classical, neighborhood, and fuzzy–rough sets, quick reduct (Max-Dependency and rough sets) [8], fuzzy–rough quick reduct (Max-Dependency and fuzzy–rough sets) [11], neighborhood quick reduct (Max-Dependency and neighborhood rough sets) [19], minimal-redundancy maximal-relevance (mRMR) framework [25], fuzzy–rough-set-based mRMR framework (fuzzy–rough mRMR) [17], and PCA [2]. All the algorithms are implemented in C language and run in Ubuntu 11.04 environment with 64-b support having machine configurations of Pentium Core 2 Quad processor at 2.66 GHz, 4-MB L2 cache, and 4-GB DDR2 RAM.

A. Class Prediction Methods

The following three pattern classifiers are used to evaluate the performances of different dimensionality reduction methods with respect to several real-life data sets.

1) SVM: The SVM [26] 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 the present work, linear kernels are used.

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

3) C4.5 Decision Tree: The C4.5 [27] is a popular decision-tree-based classification algorithm. It is used for evaluating the effectiveness of the reduced feature set for classification. The selected feature set is fed to the C4.5 for building classification models. The C4.5 is used here because it performs feature selection in the process of training, and the classification models that it builds are represented in the form of decision trees, which can be further examined.

B. Description of Data Sets

To evaluate the performances of different dimensionality reduction methods, several benchmark data sets such as Satimage, Segmentation, Isolet, and Multiple Features data sets of UCI Machine Learning Repository [28] and Breast Cancer I, Colon Cancer, Lung Cancer, Leukemia I, Breast Cancer II, and Leukemia II of Kent Ridge Bio-medical Data Set Repository [29] are used. A brief description of these data sets is also available in [21].

To compute the classification accuracy of the C4.5, K-NN rule, and SVM, both ten-fold cross-validation (CV) and training–testing are performed. The ten-fold CV is performed on the Breast Cancer I, Colon Cancer, Lung Cancer, Leukemia I, Isolet, and Multiple Features data sets, while the training–testing is done on the Satimage, Breast Cancer II, Leukemia II, and Segmentation data sets.

C. Optimum Values of Different Parameters

The multiplicative parameter η in (27) of fuzzy–rough sets controls the extent of overlapping between the 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 $\bar{c}_{\text{medium}}$ such that, ultimately, there is insignificant overlapping between the π functions low and medium or medium and high. On the other hand, as η is increased, the radius σ_{medium} increases around $\bar{c}_{\text{medium}}$ so that the amount of overlapping between the π functions increases.

The parameter r in (20) is the moment order of the significance values of all features present at a particular iteration. In effect, it controls the size of the insignificant, dispensable, and significant feature sets corresponding to the candidate feature of that iteration. Hence, the quality of the extracted features at various iterations, as well as the overall performance of the proposed dimensionality reduction method, very much depends on the value of r. Let $S = \{r, \eta\}$ be the set of parameters and $S^* = \{r^\star, \eta^\star\}$ be the set of optimal parameters. To find out the optimum values of r and η for a given data set, the class separability index [1] is used.

The class separability index $\mathcal{S}$ of a data set is defined as $\mathcal{S} = \text{trace}(V_W^{-1} V_B)$, where V_W is the within-class scatter matrix and V_B is the between-class scatter matrix, defined as follows:

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

$$V_B = \sum_{j=1}^{C} \pi_j (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^{\mathrm{T}}$$

$$\bar{\mu} = E\{X\} = \sum_{j=1}^{C} \pi_j \mu_j$$

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

For all data sets, the values of moment order r and multiplicative parameter η vary from one to five and from 0.6 to 1.5, respectively. Fig. 1 shows the variation of class separability index $\mathcal{S}$ for fuzzy–rough sets with respect to different values of r and η. The results are reported for the training samples of the Isolet, Multiple Features, and Segmentation data sets. From the results shown in Fig. 1, it is seen that, as the values of both r and η increase, the class separability index $\mathcal{S}$ also increases and attains its maximum value at the particular values of $r^\star$ and $\eta^\star$. Hence, the optimum values of $r^\star$ and $\eta^\star$ are obtained using the following relation:

$$S^\star = \arg\max_{S} \{\mathcal{S}\}. \tag{29}$$

Table I presents the optimum values of r and η for fuzzy–rough sets obtained using (29), along with the optimum values of r for classical rough sets and $\{r, \Phi\}$ for neighborhood rough sets, on different data sets. From the results reported in Table I, it can be seen that the optimum value of r is either two or three, while that of η varies in between 0.8 and 1.5, in the case of the proposed fuzzy–rough dimensionality reduction method.

D. Effectiveness of Parameter Optimization Technique

In order to establish the effectiveness of the proposed method for finding the optimum values of different parameters, extensive experimentation is done on different real-life data sets. To compute the relevance and significance of the feature set in the proposed dimensionality reduction method, classical or Pawlak's, neighborhood, and fuzzy–rough sets are used.

Table II presents the performances of the classical, neighborhood, and fuzzy–rough sets on the Satimage, Breast II, Leukemia II, and Segmentation data based on training–testing and the Colon Cancer, Breast Cancer I, Lung Cancer, Leukemia I, Isolet, and Multiple Features data based on ten-fold

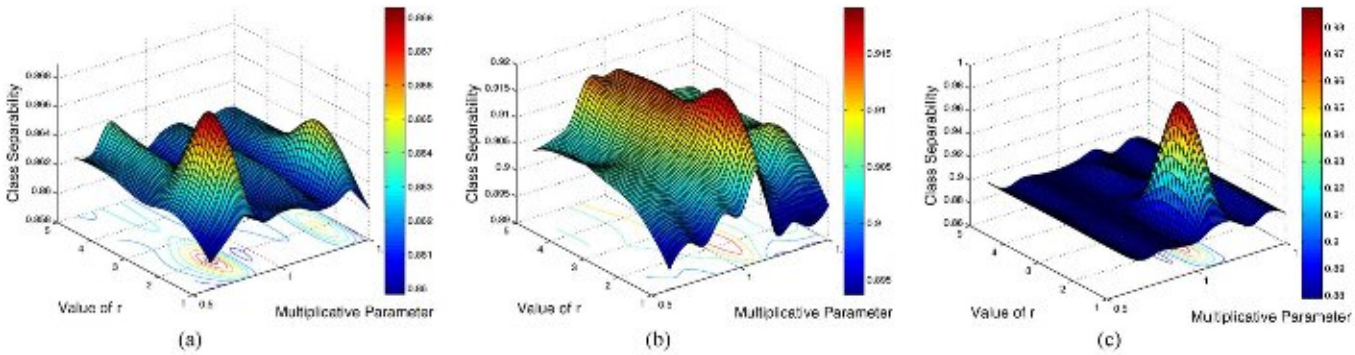

Fig. 1. Variation of the class separability index for different values of moment order r and multiplicative parameter η.

TABLE I
OPTIMUM VALUES OF THE PARAMETERS FOR THE CLASSICAL, NEIGHBORHOOD, AND FUZZY–ROUGH SETS ON DIFFERENT DATA SETS

Experimental Setup	Different Data Sets	Different Rough Sets		
		Classical: $S^* = r^*$	Neighborhood: $S^* = \{r^*, \Phi^*\}$	Fuzzy: $S^* = \{r^*, \eta^*\}$
10-fold Cross Validation	Colon Cancer	3	{4, 0.25}	{2, 1.0}
	Breast Cancer I	3	{5, 0.30/0.35}	{2, 1.1}
	Lung Cancer	3	{3/4, 0.30/0.35}	{2, 1.5}/{3, 1.3/1.5}/{4, 1.4/1.5}
	Leukemia I	3	{3, 0.10/0.20}/{4, 0.10}	{3, 0.8/1.0}
	Isolet	2	{4, 0.30}	{2, 0.8}
	Multiple Features	2	{3, 0.05/0.25}	{2, 1.1}
Training/ Testing	Satimage	2	{2, 0.40}	{3, 1.2}
	Segmentation	2	{2, 0.05}	{2, 1.2}
	Leukemia II	4	{2, 0.25}	{3, 0.8/1.3}/{3, 0.9/1.3}
	Breast Cancer II	4/5	{2, 0.20/0.25}	{3, 1.2}

CV. For ten-fold CV, only the mean of the best accuracy of ten-fold is presented. The results and subsequent discussions are analyzed in this table with respect to the classification accuracy of K-NN, SVM, and C4.5. The best test accuracy obtained from all possible parameter values on each data set is compared with the test accuracy corresponding to the best training accuracy and that for the optimum parameters.

All the results reported in Table II confirm that the test accuracy obtained using the optimum parameters is higher than the test accuracy corresponding to the best training accuracy and comparable with the best test accuracy in most of the cases, irrespective of the rough sets, classifiers, and data sets used. Out of 30 cases, the test accuracy obtained using the proposed technique is exactly the same with the best test accuracy in 22, 20, and 25 cases for the classical, neighborhood, and fuzzy–rough sets, respectively. For fuzzy–rough sets, the test accuracy corresponding to the best training accuracy is better than that of the proposed technique in only two cases, while for classical and neighborhood rough sets, it is only seven and ten cases, respectively.

E. Performances of Various Rough Set Models

In the dimensionality reduction method, the reduced feature set is always relative to a certain feature evaluation index. In general, different evaluation indices may lead to different reduced feature subsets. To establish the effectiveness of fuzzy–rough sets over Pawlak's or classical and neighborhood rough sets, extensive experiments are done on various data sets. Table III presents the comparative performances of different rough set models for a simultaneous attribute selection and feature extraction task. The results and subsequent discussions are presented in this table with respect to the classification accuracy of the K-NN, SVM, and C4.5 on test samples considering the optimum parameter values.

Both training–testing and ten-fold CV are performed to compute the classification accuracy of the SVM, C4.5, and K-NN. In the case of ten-fold CV, the means and standard deviations of the best classification accuracy obtained in different folds are computed for the Breast Cancer I, Colon Cancer, Lung Cancer, Leukemia I, Isolet, and Multiple Features data sets. Tests of significance are performed for the inequality of means (of the best classification accuracy of the SVM, C4.5, and K-NN rule) obtained using the fuzzy–rough sets and other rough sets. Since both the mean pairs and the variance pairs are unknown and different, a generalized version of t-test is used here. The aforementioned problem is the classical Behrens–Fisher problem in hypothesis testing. The test statistic, described and tabled in [30], is of the form

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\lambda_1 {\sigma_1}^2 + \lambda_2 {\sigma_2}^2}} \tag{30}$$

where μ_1 and μ_2 are the means, σ_1 and σ_2 are the standard deviations, $\lambda_1 = 1/n_1$, and $\lambda_2 = 1/n_2$, with n_1 and n_2 being the number of observations. Table III reports the individual means and standard deviations and the value of the test statistic computed. The corresponding tabled value is 1.81 at an error probability level of 0.05. If the computed value is greater than the tabled value, the means are significantly different.

From the results reported in Table III, it can be seen that the proposed dimensionality reduction method based on fuzzy–rough sets attains maximum classification accuracy of

TABLE II
CLASSIFICATION ACCURACY FOR THE OPTIMUM PARAMETERS ON DIFFERENT DATA SETS

Different Models	Different Data Sets	Test Accuracy of K-NN			Test Accuracy of SVM			Test Accuracy of C4.5		
		Best	Best Training	Proposed	Best	Best Training	Proposed	Best	Best Training	Proposed
Classical Rough Sets	Colon	96.46	96.46	95.16	96.77	96.77	95.16	100	100	100
	Breast I	100	100	100	100	100	100	100	100	100
	Lung	97.53	94.12	97.53	95.56	95.56	95.56	99.95	99.95	99.00
	Leukemia I	100	100	100	100	100	100	100	100	100
	Isolet	86.01	86.01	80.52	86.97	86.97	86.97	79.62	79.62	79.62
	Mult.Feat.	94.60	94.60	91.20	97.40	97.40	97.40	98.25	98.25	98.25
	Satimage	84.65	84.65	84.65	84.45	82.68	84.45	85.90	82.40	82.40
	Segmentation	87.61	87.61	87.61	91.38	91.38	91.38	90.90	90.90	90.90
	Leukemia II	91.07	91.07	91.07	91.07	91.07	91.07	91.07	91.07	91.07
	Breast II	89.47	89.47	84.21	94.73	94.73	89.47	100	100	100
Neighborhood Rough Sets	Colon	100	100	100	100	100	100	100	100	100
	Breast I	100	100	100	100	100	100	100	100	100
	Lung	99.50	99.50	99.50	96.06	96.06	95.56	99.50	99.50	99.15
	Leukemia I	100	100	100	100	100	100	100	100	100
	Isolet	88.86	88.86	88.86	88.64	88.64	88.14	80.52	80.52	79.62
	Mult.Feat.	97.80	97.80	97.80	97.80	97.40	97.80	98.60	98.60	97.80
	Satimage	87.50	87.50	87.50	84.75	83.60	83.60	86.00	84.65	83.95
	Segmentation	85.76	85.76	85.76	91.38	91.38	91.38	89.98	89.98	89.98
	Leukemia II	93.75	93.75	91.07	93.75	93.75	90.17	91.07	91.07	90.17
	Breast II	94.73	94.73	94.73	94.73	94.73	94.73	100	100	100
Fuzzy Rough Sets	Colon	100	100	100	100	100	98.39	100	100	100
	Breast I	100	100	100	100	100	100	100	100	100
	Lung	100	100	100	100	100	100	100	100	100
	Leukemia I	100	100	100	100	100	100	100	100	100
	Isolet	86.89	86.23	86.23	88.78	88.78	88.78	83.27	83.27	83.27
	Mult.Feat.	96.75	96.75	96.75	97.80	97.80	97.80	98.65	98.65	98.65
	Satimage	87.95	87.55	87.55	87.35	87.35	87.35	87.35	87.35	87.20
	Segmentation	86.29	86.24	86.24	92.33	92.33	92.33	90.33	90.33	90.33
	Leukemia II	94.64	94.64	94.64	93.75	93.75	93.75	93.75	93.75	93.75
	Breast II	94.73	94.73	94.73	94.73	94.73	94.73	100	100	100

TABLE III
COMPARATIVE PERFORMANCES OF VARIOUS ROUGH SET MODELS ON DIFFERENT DATA SETS

Experimental Setup	Different Data Sets	Different Statistics	Test Accuracy of K-NN			Test Accuracy of SVM			Test Accuracy of C4.5		
			Classical	Neighbor	Fuzzy	Classical	Neighbor	Fuzzy	Classical	Neighbor	Fuzzy
10-fold Cross Validation	Colon	Mean	95.16	100	100	95.16	100	98.39	100	100	100
		StdDev	0.74	0.00	0.00	0.89	0.00	0.34	0.00	0.00	0.00
		Comp	20.68	-		10.63	-14.94		-	-	
	Breast I	Mean	100	100	100	100	100	100	100	100	100
		StdDev	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		Comp	-	-		-	-		-	-	
	Lung	Mean	97.53	99.50	100	95.56	95.56	100	99.00	99.15	100
		StdDev	1.23	0.29	0.00	0.76	1.35	0.00	0.24	0.26	0.00
		Comp	6.36	5.53		18.43	10.44		9.25	7.55	
	Leukemia I	Mean	100	100	100	100	100	100	100	100	100
		StdDev	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		Comp	-	-		-	-		-	-	
	Isolet	Mean	80.52	88.86	86.23	86.97	88.14	88.78	79.62	79.62	83.27
		StdDev	2.35	2.90	1.35	2.85	3.36	1.37	2.35	2.30	1.38
		Comp	6.88	-2.60		1.81	0.57		4.23	4.20	
	Multiple Features	Mean	91.20	97.80	96.75	97.40	97.80	97.80	98.25	97.80	98.65
		StdDev	1.37	1.25	1.02	1.89	0.24	0.94	0.83	0.24	1.14
		Comp	10.29	-2.06		0.60	0.00		0.90	2.33	
Training/ Testing	Satimage	Classification Accuracy	84.65	87.50	87.55	84.45	83.60	87.35	82.40	83.95	87.20
	Segmentation		87.61	85.76	86.24	91.38	91.38	93.33	90.90	89.98	90.33
	Leukemia II		91.07	91.07	94.64	91.07	90.17	93.75	91.07	90.17	93.75
	Breast II		84.21	94.73	94.73	89.47	94.73	94.73	100	100	100

the K-NN, SVM, and C4.5 in most of the cases. Out of 12 cases of training–testing, the proposed method with fuzzy–rough sets achieves the highest classification accuracy in ten cases, while that with classical or Pawlak's rough sets attains it only in two cases. On the other hand, among the 36 comparisons of ten-fold CV, the proposed method with fuzzy–rough sets provides significantly better results in 14 cases and better results but not significantly in four cases, while significantly better results are achieved only in three cases using neighborhood rough sets. In all other cases, the performances of different rough sets are the same. In brief, out of the total 30 cases, the classical and neighborhood rough sets attain higher classification accuracy than the fuzzy–rough sets in two and three cases, respectively. In all other cases, fuzzy–rough sets provide higher or comparable classification accuracy, irrespective of the data sets, experimental setup, and classifiers used. The better performance of the fuzzy–rough sets is achieved due to the fact that it can capture uncertainties associated with the data more accurately.

TABLE IV
COMPARATIVE PERFORMANCE ANALYSIS OF DIFFERENT METHODS ON SATIMAGE, BREAST CANCER II, LEUKEMIA II, AND SEGMENTATION

Methods/ Algorithms	Different Rough Sets	Satimage			Segmentation			Leukemia II			Breast II		
		K-NN	SVM	C4.5	K-NN	SVM	C4.5	K-NN	SVM	C4.5	K-NN	SVM	C4.5
Max-Relevance	Classical	69.20	67.25	67.45	63.16	57.89	63.16	81.25	81.25	81.25	67.90	78.86	77.48
	Neighborhood	77.70	74.65	79.60	78.90	73.68	73.68	83.04	80.36	80.36	76.81	78.86	77.48
	Fuzzy	77.90	76.40	80.00	78.90	78.90	78.90	83.04	82.14	82.14	82.10	80.86	78.76
Max-Dependency	Classical	70.40	67.30	67.45	63.16	57.89	68.42	82.14	82.14	82.14	67.90	82.48	81.78
	Neighborhood	83.20	81.70	81.45	78.95	73.68	73.68	85.71	83.03	83.03	81.45	82.48	82.80
	Fuzzy	83.20	83.00	82.80	78.90	78.90	78.90	85.71	84.82	84.82	84.19	83.76	82.76
MRMS	Classical	74.00	73.85	74.10	72.67	74.10	74.67	84.82	85.71	85.71	84.21	84.21	84.21
	Neighborhood	83.40	85.00	85.15	80.52	82.57	83.24	87.50	89.29	88.39	84.21	89.47	89.47
	Fuzzy	84.10	84.10	84.10	80.76	83.95	85.14	88.39	90.18	89.29	89.47	94.74	94.74
Classical mRMR		75.45	75.40	75.35	72.81	73.76	74.33	84.82	84.82	84.82	84.21	84.21	89.47
Fuzzy-Rough mRMR		83.95	84.60	83.70	80.33	84.10	84.71	87.50	89.29	90.18	89.47	89.47	94.74
PCA		82.55	83.95	82.00	78.94	89.47	94.73	80.35	78.59	79.46	77.30	79.50	74.10
Proposed Fuzzy-Rough		87.55	87.35	87.20	86.24	92.33	90.33	94.64	93.75	93.75	94.73	94.73	100

TABLE V
COMPARATIVE PERFORMANCE ANALYSIS OF DIFFERENT METHODS ON COLON CANCER, BREAST CANCER I, AND LUNG CANCER

Different Classifiers	Methods/ Algorithms	Different Rough Sets	Colon			Breast I			Lung		
			Mean	StdDev	Comp	Mean	StdDev	Comp	Mean	StdDev	Comp
K-NN	Max-Relevance	Classical	85.48	1.93	14.82	81.63	2.34	24.85	73.20	2.93	24.19
		Neighborhood	85.48	2.31	19.86	83.67	2.55	20.24	81.23	2.43	23.64
		Fuzzy	85.48	2.31	19.86	83.67	2.09	24.70	82.32	2.05	27.27
	Max-Dependency	Classical	87.09	3.40	7.33	85.71	3.23	13.98	73.20	3.12	22.96
		Neighborhood	87.09	3.40	11.99	88.67	2.38	15.07	88.66	1.62	20.87
		Fuzzy	88.71	2.78	12.85	88.67	2.44	14.71	87.19	1.64	24.73
	MRMS	Classical	87.09	2.11	19.35	88.67	2.83	12.66	79.31	3.19	20.51
		Neighborhood	88.71	2.62	13.63	88.67	2.69	13.32	88.18	2.22	16.85
		Fuzzy	88.71	2.58	13.84	88.67	2.06	17.39	88.67	2.73	13.14
	Fuzzy-Rough mRMR		87.10	1.57	26.00	89.80	0.45	71.75	85.22	1.43	32.68
	Classical mRMR		87.09	1.88	21.72	88.67	1.28	27.99	81.77	1.48	38.95
	PCA		86.84	1.30	31.99	89.79	1.30	21.86	86.69	1.23	34.25
	Proposed Fuzzy-Rough		100	0.00		100	0.00		100	0.00	
SVM	Max-Relevance	Classical	64.52	2.62	35.01	79.59	2.10	30.69	76.28	3.02	19.61
		Neighborhood	85.48	1.96	23.43	89.80	2.32	13.92	83.98	2.26	13.94
		Fuzzy	85.48	1.67	23.93	87.76	2.50	15.50	86.74	2.32	18.07
	Max-Dependency	Classical	64.51	3.50	26.84	83.45	1.22	43.00	76.28	3.42	17.41
		Neighborhood	88.71	3.11	11.47	93.87	1.35	14.39	87.19	1.30	14.17
		Fuzzy	88.71	2.44	12.44	91.28	1.44	19.18	87.19	1.54	26.24
	MRMS	Classical	77.42	3.16	20.86	87.76	3.13	12.42	81.28	2.66	22.25
		Neighborhood	83.87	2.52	18.04	91.84	1.03	25.08	87.68	2.82	13.82
		Fuzzy	85.48	2.78	14.56	93.88	1.37	14.15	88.18	3.01	12.43
	Fuzzy-Rough mRMR		85.48	1.62	24.63	92.00	1.22	20.74	86.21	2.13	20.49
	Classical mRMR		77.42	2.47	26.59	87.76	1.52	25.49	81.28	1.28	46.25
	PCA		88.70	1.25	23.67	93.87	1.88	10.33	87.17	1.24	32.61
	Proposed Fuzzy-Rough		98.39	0.34		100	0.00		100	0.00	
C4.5	Max-Relevance	Classical	64.50	2.68	41.94	81.63	2.43	23.91	68.50	3.27	29.34
		Neighborhood	87.10	2.77	14.74	89.80	2.65	12.18	88.40	2.41	13.97
		Fuzzy	88.71	1.89	18.91	89.80	2.65	12.18	87.29	2.37	16.96
	Max-Dependency	Classical	64.50	2.68	41.94	83.45	1.22	43.00	68.50	3.59	26.77
		Neighborhood	88.70	1.02	35.17	91.80	1.56	16.67	89.06	1.17	26.16
		Fuzzy	90.32	1.22	25.07	91.84	1.48	17.46	88.18	1.81	20.71
	MRMS	Classical	83.87	3.31	15.41	85.71	3.04	14.86	72.91	2.18	39.31
		Neighborhood	90.32	1.48	20.68	91.84	1.44	17.94	89.16	2.55	13.44
		Fuzzy	90.32	1.48	20.68	93.88	1.83	10.59	88.67	2.73	13.14
	Fuzzy-Rough mRMR		90.32	1.43	21.41	91.84	1.16	22.27	89.16	1.25	27.42
	Classical mRMR		82.26	1.27	44.20	85.71	2.33	19.39	77.83	1.62	43.28
	PCA		88.70	1.11	32.25	91.83	1.78	14.51	89.16	1.43	23.90
	Proposed Fuzzy-Rough		100	0.00		100	0.00		100	0.00	

F. Performances of Different Algorithms

Finally, Tables IV–VII compare the performance of the proposed fuzzy–rough simultaneous feature selection and extraction algorithm with those of different existing feature selection and extraction algorithms on various data sets.

From the results reported in Table IV, it is seen that the proposed dimensionality reduction method achieves the highest classification accuracy of SVM, C4.5, and K-NN in 11 cases out of the total 12 cases, while PCA attains the highest classification accuracy in only one case. The proposed method also provides higher classification accuracy than the Max-Relevance, Max-Dependency, and MRMS criteria in all cases, irrespective of the classifiers, rough sets, and data sets used. Tables V and VI report the performances of different methods in the case of ten-fold CV, along with the results of the test of significance, for the K-NN, SVM, and C4.5. From the results

TABLE VI
COMPARATIVE PERFORMANCE ANALYSIS OF DIFFERENT METHODS ON LEUKEMIA I, ISOLET, AND MULTIPLE FEATURES

Different Classifiers	Methods/ Algorithms	Different Rough Sets	Leukemia I			Isolet			Multiple Features		
			Mean	StdDev	Comp	Mean	StdDev	Comp	Mean	StdDev	Comp
K-NN	Max-Relevance	Classical	84.72	1.27	37.96	58.68	2.14	22.22	84.55	1.39	10.79
		Neighborhood	88.89	1.39	25.26	74.27	2.74	11.56	83.20	2.53	16.36
		Fuzzy	87.50	2.38	16.59	78.54	2.88	7.65	85.35	2.18	14.97
	Max-Dependency	Classical	89.46	1.12	29.71	58.68	2.14	22.22	87.90	2.45	3.73
		Neighborhood	91.42	0.97	27.94	79.68	2.52	7.55	88.70	1.62	14.09
		Fuzzy	91.67	1.21	21.80	79.85	2.11	8.05	90.25	1.49	11.35
	MRMS	Classical	88.89	2.52	13.95	67.46	3.88	14.45	87.95	3.02	8.73
		Neighborhood	93.06	1.28	17.17	79.42	2.93	6.68	89.35	1.82	11.21
		Fuzzy	93.06	1.32	16.65	82.28	1.62	5.93	90.65	1.77	9.43
	Fuzzy-Rough mRMR		93.06	1.62	13.57	82.24	2.19	4.91	89.35	2.11	9.98
	Classical mRMR		88.89	1.83	19.22	79.85	1.42	10.30	87.35	1.39	17.22
	PCA		84.28	1.33	40.32	76.84	1.49	14.79	78.80	3.35	22.14
	Proposed Fuzzy-Rough		100	0.00		86.23	1.35		96.75	1.02	
SVM	Max-Relevance	Classical	87.50	2.42	16.33	58.56	2.95	21.91	82.25	2.10	16.95
		Neighborhood	93.06	1.55	14.18	75.67	2.66	9.38	83.75	2.02	21.85
		Fuzzy	88.89	1.38	25.48	78.37	2.48	11.62	85.45	1.92	18.30
	Max-Dependency	Classical	89.24	1.29	26.30	58.56	1.34	28.55	82.25	1.30	20.85
		Neighborhood	98.61	0.35	12.67	78.68	3.13	6.63	85.46	2.25	17.26
		Fuzzy	95.83	0.83	15.89	79.54	2.44	10.44	85.65	1.98	17.54
	MRMS	Classical	90.28	1.05	29.30	74.67	3.92	10.75	84.10	3.85	10.94
		Neighborhood	97.22	0.40	21.98	80.87	2.11	9.95	89.55	2.73	9.04
		Fuzzy	97.22	0.45	19.54	83.12	2.63	6.04	92.85	2.55	5.76
	Fuzzy-Rough mRMR		95.89	1.12	11.60	82.52	2.48	6.99	91.15	1.39	12.56
	Classical mRMR		90.28	1.52	20.24	74.63	1.73	20.29	83.35	1.83	22.25
	PCA		84.88	1.49	32.03	81.28	1.83	10.39	82.51	0.48	46.04
	Proposed Fuzzy-Rough		100	0.00		88.78	1.37		97.80	0.93	
C4.5	Max-Relevance	Classical	88.89	2.91	12.08	58.44	1.87	22.29	84.15	1.89	21.61
		Neighborhood	88.89	2.73	12.88	64.89	1.68	16.33	85.50	2.45	15.80
		Fuzzy	90.28	1.68	18.31	65.48	2.12	22.24	85.35	2.54	15.25
	Max-Dependency	Classical	94.20	2.11	8.68	58.44	1.87	22.29	84.20	2.11	19.57
		Neighborhood	95.56	0.79	17.82	64.89	1.68	16.33	89.55	1.64	15.70
		Fuzzy	95.83	1.01	13.03	65.48	2.12	22.24	86.65	1.70	18.59
	MRMS	Classical	93.06	1.16	18.95	67.57	3.66	12.69	85.05	2.73	14.56
		Neighborhood	95.83	1.35	9.77	72.35	2.83	10.96	89.25	3.22	8.71
		Fuzzy	97.22	0.45	19.54	75.48	3.84	6.04	86.95	3.72	9.52
	Fuzzy-Rough mRMR		95.83	1.18	11.18	74.44	2.61	9.45	89.45	1.38	16.32
	Classical mRMR		93.06	1.24	17.72	71.68	1.87	15.75	84.85	1.44	23.86
	PCA		85.46	1.79	35.66	66.89	2.57	17.77	84.60	0.39	37.23
	Proposed Fuzzy-Rough		100	0.00		83.27	1.38		98.65	1.13	

TABLE VII
EXECUTION TIME (IN SECONDS) OF DIFFERENT METHODS FOR VARIOUS DATA SETS

Methods/ Algorithms	Different Rough Sets	Different Benchmark Data Sets									
		Colon	Breast I	Lung	Leukem I	Isolet	Mult.Feat.	Satimage	Segmnt	Leukem II	Breast II
Max-Relevance	Classical	0.2	0.3	2.0E1	0.8	8.3	4.3	0.1	0.1	5.2	3.0E1
	Neighborhood	5.4E2	6.3E2	8.1E4	7.3E2	2.1E4	1.0E4	8.4E2	6.3E2	5.7E4	1.5E5
	Fuzzy	0.2	2.3	8.1	2.1	2.3E1	7.1	1.7E1	1.6E1	3.2E1	4.5E1
Max-Dependency	Classical	2.6E1	1.4E1	8.3E1	1.8E1	1.3E2	4.6E1	4.9	3.3	9.4E2	3.3E2
	Neighborhood	1.2E4	9.3E3	6.2E5	1.1E4	2.2E5	6.7E4	2.4E4	2.8E4	5.3E5	1.8E6
	Fuzzy	9.8E3	9.8E3	5.9E5	9.4E3	2.6E5	5.2E4	3.3E4	2.7E4	5.5E5	1.3E6
MRMS	Classical	0.2	0.4	4.3E1	1.5	9.4	5.9	0.3	0.2	7.9	3.8E1
	Neighborhood	7.3E2	8.7E2	9.7E4	9.3E2	3.0E4	2.1E4	1.0E3	9.5E2	8.7E4	2.8E5
	Fuzzy	0.4	2.8	9.5	2.7	3.2E1	1.0E1	4.3E1	2.3E1	2.7E1	4.8E1
Classical mRMR		0.5	5.5	1.8E1	5.4	7.4	0.4	0.1	0.1	5.0	5.4
Fuzzy-Rough mRMR		1.7	6.1E1	3.8E1	3.1E1	5.6E1	1.4E1	1.2	1.8	6.5	5.6
PCA		2.2E1	3.0E1	3.2E1	3.8E1	1.1E1	1.8E1	7.8E3	8.0E2	4.0E3	4.0E1
Proposed Fuzzy-Rough		6.4	4.3E1	4.5E1	8.2E1	4.9E1	1.0E1	2.8E1	3.1E1	6.3E1	7.3E1

reported in these tables, it can be seen that the proposed method attains significantly better accuracy than the other algorithms in all cases, irrespective of the data sets and classifiers used. Moreover, Table VII reports the execution time of different algorithms. The significantly lesser time of the proposed algorithm is achieved due to its low computational complexity.

Hence, all the results reported in Tables IV–VI confirm that the proposed fuzzy–rough dimensionality reduction method selects a set of features having the highest classification accuracy of the K-NN, SVM, and C4.5 in most of the cases, irrespective of the data sets. Also, the proposed method can potentially yield significantly better results than the existing algorithms. The better performance of the proposed method is achieved due to the fact that it provides an efficient way to simultaneously select and extract features for classification. In effect, a reduced set of features having maximum relevance and significance is being obtained using the proposed method.

V. Conclusion

One of the important problems in pattern recognition and data mining, particularly given the explosive growth of available information, is the dimensionality reduction using feature selection and/or feature extraction. In this regard, this paper has presented a novel dimensionality reduction method, integrating judiciously the theory of fuzzy–rough sets and the merits of both attribute selection and feature extraction. An efficient algorithm has been introduced by performing simultaneous feature selection and extraction. It uses the concept of fuzzy–rough feature significance for finding significant and relevant features of real-valued data sets. Finally, the effectiveness of the proposed method has been presented, along with a comparison with other related algorithms, on real-life data sets. This formulation is geared toward maximizing the utility of fuzzy–rough sets, feature selection, and feature extraction with respect to knowledge discovery tasks. Through these investigations and experiments, the potential utility of fuzzy–rough sets for dimensionality reduction is demonstrated.

References

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

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification and Scene Analysis*. Hoboken, NJ: Wiley, 2000.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. 7/8, pp. 1157–1182, Mar. 2003.

[4] H.-L. Wei and S. Billings, "Feature subset selection and ranking for data dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 162–166, Jan. 2007.

[5] M. Bressan and J. Vitria, "On the selection and classification of independent features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1312–1317, Oct. 2003.

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

[7] P. Maji and S. K. Pal, *Rough–Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging*. Hoboken, New Jersey: IEEE Press, 2012.

[8] A. Chouchoulas and Q. Shen, "Rough set-aided keyword reduction for text categorization," *Appl. Artif. Intell.*, vol. 15, no. 9, pp. 843–873, Oct. 2001.

[9] P. Maji and S. Paul, "Rough set based maximum relevance–maximum significance criterion and gene selection from microarray data," *Int. J. Approx. Reason.*, vol. 52, no. 3, pp. 408–426, Mar. 2011.

[10] N. Parthalain, Q. Shen, and R. Jensen, "A distance measure approach to exploring the rough set boundary region for attribute reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 3, pp. 305–317, Mar. 2010.

[11] 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.

[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] S. Zhao, E. C. C. Tsang, D. Chen, and X. Wang, "Building a rule-based classifier: A fuzzy–rough set approach," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 5, pp. 624–638, May 2010.

[14] 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.

[15] E. C. C. Tsang, D. Chen, D. S. Yeung, X.-Z. Wang, and J. Lee, "Attributes reduction using fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 5, pp. 1130–1141, Oct. 2008.

[16] R. Jensen and Q. Shen, "New approaches to fuzzy–rough feature selection," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 4, pp. 824–838, Aug. 2009.

[17] 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.

[18] D. Chen, L. Zhang, S. Zhao, Q. Hu, and P. Zhu, "A novel algorithm for finding reducts with fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 2, pp. 385–389, Apr. 2012.

[19] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection," *Inf. Sci.*, vol. 178, no. 18, pp. 3577–3594, Sep. 2008.

[20] Q. Hu, W. Pedrycz, D. Yu, and J. Lang, "Selecting discrete and continuous features based on neighborhood decision error minimization," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 1, pp. 137–150, Feb. 2010.

[21] P. Maji and P. Garai, "On fuzzy–rough attribute selection: Criteria of max-dependency, max-relevance, min-redundancy, and max-significance," *Appl. Soft Comput.*, to be published.

[22] 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.

[23] P. Maji, "Fuzzy–rough supervised attribute clustering algorithm and classification of microarray data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 222–233, Feb. 2011.

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

[25] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

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

[27] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[28] A. Frank and A. Asuncion, UCI Machine Learning Repository, Univ. California, Irvine, CA. [Online]. Available: http://archive.ics.uci.edu/ml

[29] Kent Ridge Bio-medical Data Set Repository. [Online]. Available: http://sdmc.lit.org.sg/GEDatasets/Datasets.html

[30] A. Aspin, "Tables for use in comparisons whose accuracy involves two variances, separately estimated," *Biometrika*, vol. 36, no. 3/4, pp. 245–271, Dec. 1949.

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 published around 80 papers in international journals and conferences. He is an author of a book published by Wiley–IEEE Computer Society Press and also a Reviewer of many international journals. His research interests include pattern recognition, machine learning, soft computing, computational biology and bioinformatics, and medical image processing.

Dr. Maji was the recipient of the 2008 Microsoft Young Faculty Award from Microsoft Research Laboratory India Pvt., the 2009 Young Scientist Award from the National Academy of Sciences, India, and the 2011 Young Scientist Award from the Indian National Science Academy and was selected as the 2009 Young Associate of the Indian Academy of Sciences.

Partha Garai received the B.Sc. degree in physics from the University of Calcutta, Kolkata, India, in 2002 and the M.C.A. degree and the M.Tech. degree in information technology from the West Bengal University of Technology, Kolkata, in 2006 and 2009, respectively.

He is currently a Research Scholar with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata. He has published a few papers in international journals and conferences. His research interests include pattern recognition, machine learning, soft computing, and very large scale integration.

Mr. Garai has received the INSPIRE fellowship of the Department of Science and Technology, Government of India, for the years 2011–2016.