Applied Soft Computing xxx (2012) xxx–xxx

On fuzzy-rough attribute selection: Criteria of Max-Dependency, Max-Relevance, Min-Redundancy, and Max-Significance

Pradipta Maji*, Partha Garai

Machine Intelligence Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, West Bengal, India

Keywords:
Pattern recognition
Data mining
Attribute selection
Rough sets
Classification

A B S T R A C T

Attribute selection is one of the important problems encountered in pattern recognition, machine learning, data mining, and bioinformatics. It refers to the problem of selecting those input attributes or features that are most effective to predict the sample categories. In this regard, rough set theory has been shown to be successful for selecting relevant and nonredundant attributes from a given data set. However, the classical rough sets are unable to handle real valued noisy features. This problem can be addressed by the fuzzy-rough sets, which are the generalization of classical rough sets. A feature selection method is presented here based on fuzzy-rough sets by maximizing both relevance and significance of the selected features. This paper also presents different feature evaluation criteria such as dependency, relevance, redundancy, and significance for attribute selection task using fuzzy-rough sets. The performance of different rough set models is compared with that of some existing feature evaluation indices based on the predictive accuracy of nearest neighbor rule, support vector machine, and decision tree. The effectiveness of the fuzzy-rough set based attribute selection method, along with a comparison with existing feature evaluation indices and different rough set models, is demonstrated on a set of benchmark and microarray gene expression data sets.

1. Introduction

Attribute or feature selection 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 $d < m$. The main objective of this task is to retain 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 classification [1–3].

The problem of attribute selection has two aspects, namely, formulation of a suitable criterion to evaluate the goodness of a feature set and searching the optimal set in terms of the criterion [4]. In general, those features are considered to have optimal saliencies for which interclass (respectively, intraclass) distances are maximized (respectively, 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 [5,6].

The conventional feature selection is based on the minimal classification error, which usually requires the maximal statistical dependency of the sample categories or class labels on the data distribution in the reduced feature space. This scheme is called maximal dependency or Max-Dependency, in which, the task of feature selection is to find a feature subset from the whole feature set, which jointly have the largest dependency on the target class [7–9]. However, the main drawback of this approach is the slow computational speed. Also, the joint dependency of the features for high dimensional real life data sets cannot be estimated correctly [10,11]. Hence, although Max-Dependency feature selection might be useful to select a very small number of features, it is not appropriate for real life applications where the aim is to achieve high classification accuracy with a reasonably compact set of features.

As Max-Dependency criterion is hard to implement, an alternative is to select features based on maximal relevance or Max-Relevance criterion. Max-Relevance is to search a set of features that approximates Max-Dependency criterion with the mean value of all dependency values between individual feature and target class label. However, Max-Relevance criterion does not consider the joint effect of features on the target class. Moreover, it is likely that features selected according to Max-Relevance could have rich redundancy, that is, the dependency among these features could be large [12,13].

Some feature selection methods have been reported to reduce redundancy among the selected features directly based on minimal redundancy or Min-Redundancy criterion [14,15] or indirectly based on maximal significance or Max-Significance criterion

* Corresponding author. Tel.: +91 33 2575 3113; fax: +91 33 2578 3357.
E-mail addresses: pmaji@isical.ac.in (P. Maji), parthagarai_r@isical.ac.in (P. Garai).

[16–18]. Min-Redundancy criterion has also been studied in principal component analysis (PCA) and independent component analysis (ICA) [2,19], which aims to find nonredundant features in a transformed domain. Combining redundancy or significance criterion with relevance criterion, minimal redundancy-maximal relevance (mRMR) [10,11,20] and maximal relevance-maximal significance (MRMS) [16–18] criteria have been proposed to select relevant and nonredundant or significant features.

An optimal feature subset selected by a feature selection algorithm is always relative to a certain feature evaluation index. In general, different indices may lead to different optimal feature subsets [5,21]. However, every index tries to measure the discriminating ability of a feature or a subset of features to distinguish different class labels or sample categories. To compute the effectiveness of a feature or a subset of features, different statistical measures, Euclidean distance [10], mutual information [7–9], class separability index [1], Davies–Bouldin index [22], Dunn index [23], and fuzzy feature evaluation index [24] are widely used. 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 [25] 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 [25–34].

Rough set theory can be used to find a subset of informative features from the original attributes of a given data set with discretized attribute values [17,35]. While the quick reduct algorithm of Chouchoulas and Shen [35] is based on the principle of Max-Dependency criterion, the MRMS criterion is used in [17] for attribute selection task. However, there are usually real valued data and fuzzy information in real world applications. In rough set theory, the real valued features are divided into several discrete partitions and the dependency or quality of approximation of a feature is calculated. The inherent error that exists in discretization process is of major concern in the computation of the dependency of real valued features. Combining fuzzy and rough sets provides an important direction in reasoning with uncertainty for real valued data sets [36,37]. They are complementary in some aspects. The generalized theories of rough-fuzzy computing have been applied successfully to feature selection of real valued data [20,31,36,38]. Also, neighborhood rough sets [39] are found to be suitable for both numerical and categorical data sets. The fuzzy-rough quick reduct algorithm of Jensen and Shen [36] and neighborhood rough set based feature selection algorithm of Hu et al. [39] are based on Max-Dependency criterion, while the feature selection method based on f-information measures in fuzzy approximation spaces of Maji and Pal [20] uses the mRMR criterion.

In this regard, a fuzzy-rough feature selection method is presented, integrating judiciously the merits of fuzzy-rough sets and MRMS criterion, to provide a means by which real valued noisy features can be effectively reduced without the need for user-specified information. The proposed method selects a subset of features or condition attributes from the whole feature set by maximizing the relevance and significance of the selected features. Both relevance and significance of the features are computed using the concept of fuzzy positive regions of fuzzy-rough sets. Hence, the only information required in the proposed feature selection method is in the form of fuzzy partitions or information granules for each condition attribute. The π function in the one dimensional form is used to generate fuzzy information granules corresponding to each condition attribute, where the centers and radii of the π functions can be determined automatically from the distribution of training patterns. The fuzzy positive regions of decision attributes or class labels are computed based on the concept of fuzzy equivalence partition matrix [20]. The method can be applied to regression as well as classification problems with continuous decision attributes. The effectiveness of the proposed fuzzy-rough attribute selection method, along with a comparison with other methods, is demonstrated on a set of benchmark and microarray gene expression data sets using the predictive accuracy of nearest neighbor rule, support vector machine, and decision tree.

The structure of this paper is as follows: Section 2 briefly introduces the basic notions of rough sets, neighborhood rough sets, and fuzzy-rough sets, along with some existing feature evaluation indices. The proposed fuzzy-rough attribute selection method is described in Section 3. A few case studies and a comparison with other methods are presented in Section 4. Concluding remarks are given in Section 5.

2. Feature evaluation indices and rough sets

This section presents some existing feature evaluation indices and various rough set models for feature selection task.

2.1. Existing feature evaluation indices

Following four feature evaluation indices, namely, class separability index [1], Davies–Bouldin index [22], Dunn index [23], and fuzzy feature evaluation index [24], are considered to measure the effectiveness of the feature subset.

2.1.1. Class separability index

The class separability index $\mathcal{S}$ [1] of a data set is defined as

$$\mathcal{S} = \text{trace}(V_B^{-1} V_W), \tag{1}$$

where V_W is the within class scatter matrix and V_B is the between class scatter matrix, defined as follows:

$$V_W = \sum_{j=1}^{c} \pi_j E\{(X - v_j)(X - v_j)^T | \beta_j\} = \sum_{j=1}^{c} \pi_j \Sigma_j; \tag{2}$$

$$V_B = \sum_{j=1}^{c} \pi_j (v_j - \overline{v})(v_j - \overline{v})^T; \tag{3}$$

and

$$\overline{v} = E\{X\} = \sum_{j=1}^{c} \pi_j v_j; \tag{4}$$

where c is the number of classes, π_j is a priori probability that a pattern belongs to class β_j, X is a feature vector, $\overline{v}$ is the sample mean vector for the entire data points, v_j and Σ_j represent the sample mean and covariance matrix of class β_j, respectively, and $E\{\cdot\}$ is the expectation operator. A lower value of class separability index $\mathcal{S}$ ensures that classes are well separated by their scatter means. Hence, a good feature subset should have the value of $\mathcal{S}$ index as low as possible.

2.1.2. Davies–Bouldin index

The Davies–Bouldin (DB) index [22] is a function of the ratio of sum of within class distance to between class separation and is given by

$$\text{DB} = \frac{1}{c} \sum_{i=1}^{c} \max_{i \neq k} \left\{ \frac{S(v_i) + S(v_k)}{d(v_i, v_k)} \right\} \quad \text{for } 1 \leq i, k \leq c. \tag{5}$$

The DB index minimizes the within class distance $S(v_i)$ and maximizes the between class separation $d(v_i, v_k)$, where v_i is the sample mean of class β_i. Therefore, for a given data set and c value, the higher the similarity values within the class and the between class

separation, the lower would be the DB index value. A good feature set should have the value of DB index as low as possible.

2.1.3. Dunn index

Dunn's index [23] is also designed to identify sets of clusters that are compact and well separated. Dunn's (D) index maximizes

$$\mathrm{D} = \min_{i}\left\{\min_{i\neq k}\left\{\frac{d(v_i, v_k)}{\max_l S(v_l)}\right\}\right\} \quad \text{for } 1 \le i, k, l \le c. \tag{6}$$

A good feature subset should have the value of Dunn index as high as possible.

2.1.4. Fuzzy feature evaluation index

The fuzzy feature evaluation index is defined as [24]

$$\mathrm{FFEI} = \frac{2}{n(n-1)}\sum_{i}^{n}\sum_{j\neq i}^{n}\frac{1}{2}\left[\mu_{ij}^{R}(1-\mu_{ij}^{O}) + \mu_{ij}^{O}(1-\mu_{ij}^{R})\right], \tag{7}$$

where μ_{ij}^{O} and μ_{ij}^{R} are the degrees that both patterns x_i and x_j belong to the same cluster in the original feature space Ω_O and reduced feature space Ω_R, respectively, and n is the total number of samples. The membership function μ_{ij} can be defined as

$$\mu_{ij} = \begin{cases} 1 - \dfrac{d_{ij}}{D} & \text{if } d_{ij} \le D \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where d_{ij} is the distance between patterns x_i and x_j, and D may be expressed as

$$D = \beta d_{max}, \tag{9}$$

where d_{max} is the maximum separation between patterns in the respective feature spaces and β is an user defined constant ranging 0 to 1. In the present work, the value of β is set to 0.2. d_{ij} can be defined in many ways, like Euclidean distance. The value of FFEI decreases as the intercluster (respectively, intracluster) distances increase (respectively, decrease). Hence, the lower the value of FFEI, the more crisp is the cluster structure.

2.2. Various rough set models

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

2.2.1. Rough sets

The rough set theory begins with the notion of an approximation space, which is a pair $\langle \mathbb{U}, \mathbb{A}\rangle$, where $\mathbb{U}$ be a non-empty set, $\mathbb{U} = \{x_1, \ldots, x_i, \ldots, x_n\}$, 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 [25].

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

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

If $(x_i, x_j) \in IND(\mathbb{P})$, 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

$$\mathbb{U}/IND(\mathbb{P}) = \{[x_i]_{\mathbb{P}} : x_i \in \mathbb{U}\}, \tag{10}$$

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

Given an arbitrary set $X \subseteq \mathbb{U}$, in general, it may not be possible to describe X precisely in $\langle \mathbb{U}, \mathbb{A}\rangle$. One may characterize X by a pair of lower and upper approximations defined as follows [25]:

$$\underline{\mathbb{P}}(X) = \bigcup\{[x_i]_{\mathbb{P}} \mid [x_i]_{\mathbb{P}} \subseteq X\} \quad \text{and} \quad \overline{\mathbb{P}}(X) = \bigcup\{[x_i]_{\mathbb{P}} \mid [x_i]_{\mathbb{P}} \cap X \neq \emptyset\}. \tag{11}$$

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 non-empty 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 (respectively, upper) approximation $\underline{\mathbb{P}}(X)$ (respectively, $\overline{\mathbb{P}}(X)$) is interpreted as the collection of those elements of $\mathbb{U}$ that definitely (respectively, possibly) belong to X. The lower approximation is also called positive region sometimes, denoted as $POS_{\mathbb{P}}(X)$. A set X is said to be definable in the approximation space $\langle \mathbb{U}, \mathbb{A}\rangle$ 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 dependency between $\mathbb{C}$ and $\mathbb{D}$ can be defined as [25]:

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|POS_{\mathbb{C}}(\mathbb{D})|}{|\mathbb{U}|}, \tag{12}$$

where $POS_{\mathbb{C}}(\mathbb{D}) = \bigcup \underline{\mathbb{C}}X_i$, X_i is the ith equivalence class induced by $\mathbb{D}$ and $|\cdot|$ denotes the cardinality of a set. If $\gamma_{\mathbb{C}}(\mathbb{D}) = 1$, $\mathbb{D}$ depends totally on $\mathbb{C}$, if $0 < \gamma_{\mathbb{C}}(\mathbb{D}) < 1$, $\mathbb{D}$ depends partially on $\mathbb{C}$, and if $\gamma_{\mathbb{C}}(\mathbb{D}) = 0$, then $\mathbb{D}$ does not depend on $\mathbb{C}$. Given $\mathbb{C}$, $\mathbb{D}$ and an attribute $\mathcal{A} \in \mathbb{C}$, the significance of the attribute $\mathcal{A}$ is defined as [25]:

$$\sigma_{\mathbb{C}}(\mathbb{D}, \mathcal{A}) = \gamma_{\mathbb{C}}(\mathbb{D}) - \gamma_{\mathbb{C}-\{\mathcal{A}\}}(\mathbb{D}). \tag{13}$$

Hence, the change in dependency when an attribute is removed from the set of condition attributes, is a measure of the significance of the attribute. The higher the change in dependency, the more significant the attribute is. If the significance is 0, then the attribute is dispensable.

2.2.2. Neighborhood rough sets

Given an arbitrary $x_i \in \mathbb{U}$ 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 [39]

$$\Phi_{\mathbb{P}}(x_i) = \{x_j | x_j \in \mathbb{U}, \Delta^{\mathbb{P}}(x_i, x_j) \le \Phi\}, \tag{14}$$

where Δ is a distance function. $\Phi_{\mathbb{P}}(x_i)$ in (14) 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 second controls the size of 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, neighborhood rough set based algorithm selects different attribute subsets with the change of Δ function and Φ value [34].

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 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 round is not possible [34]. 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 neighborhood rough set model degenerates to rough sets [39].

2.2.3. 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 [37].

Let $\langle \mathbb{U}, \mathbb{A} \rangle$ represents a fuzzy approximation space and X is a fuzzy subset of $\mathbb{U}$. The fuzzy $\mathbb{P}$-lower and $\mathbb{P}$-upper approximations are then defined as follows [37]:

$$\mu_{\underline{\mathbb{P}}X}(F_i) = \inf_x \{\max\{(1 - \mu_{F_i}(x)), \mu_X(x)\}\} \quad \forall i, \tag{15}$$

$$\mu_{\overline{\mathbb{P}}X}(F_i) = \sup_x \{\min\{\mu_{F_i}(x), \mu_X(x)\}\} \quad \forall i, \tag{16}$$

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 [36]

$$\mu_{\underline{\mathbb{P}}X}(x) = \sup_{F_i \in \mathbb{U}/\mathbb{P}} \min\{\mu_{F_i}(x), \mu_{\underline{\mathbb{P}}X}(F_i)\}, \tag{17}$$

$$\mu_{\overline{\mathbb{P}}X}(x) = \sup_{F_i \in \mathbb{U}/\mathbb{P}} \min\{\mu_{F_i}(x), \mu_{\overline{\mathbb{P}}X}(F_i)\}. \tag{18}$$

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

$$\mu_{POS_{\mathbb{C}}(\mathbb{D})}(x) = \sup_{X \in \mathbb{U}/\mathbb{D}} \mu_{\underline{\mathbb{C}}X}(x), \tag{19}$$

where $\mathbb{A} = \mathbb{C} \cup \mathbb{D}$. Using the definition of fuzzy positive region, the dependency function can be defined as follows [36]:

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|\mu_{POS_{\mathbb{C}}(\mathbb{D})}(x)|}{|\mathbb{U}|} = \frac{1}{|\mathbb{U}|} \sum_{x \in \mathbb{U}} \mu_{POS_{\mathbb{C}}(\mathbb{D})}(x). \tag{20}$$

3. Fuzzy-rough attribute selection method

This section presents a feature selection algorithm, integrating judiciously the theory of fuzzy-rough sets and merits of the MRMS criterion.

3.1. Fuzzy-rough MRMS method

The real life high dimensional data set may contain a number of irrelevant and insignificant features. The presence of such features may lead to a reduction in the useful information. The selected feature subset should contain the features those have high relevance with the classes and high significance in the feature set. The features with high relevance are expected to be able to predict the classes of the samples. In contrast, the presence of insignificant features in the subset may degrade the prediction capability. A feature set with high relevance and high significance enhances the predictive capability. Accordingly, a measure is required that can assess the effectiveness of a feature set. In this paper, the theory of fuzzy-rough sets is used to select relevant and significant features from a data set.

Let $\mathbb{C} = \{\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots, \mathcal{A}_j, \ldots, \mathcal{A}_m\}$ be the set of m condition attributes or features of a given data set and $\mathbb{S} \subseteq \mathbb{C}$ with cardinality $d < m$ is the set of selected features. Define $\gamma_{\mathcal{A}_i}(\mathbb{D})$ as the relevance of the feature $\mathcal{A}_i$ with respect to the class labels $\mathbb{D}$ while $\sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_i)$ as the significance of the feature $\mathcal{A}_i$ with respect to the set $\{\mathcal{A}_i, \mathcal{A}_j\}$. The average relevance of all selected features is, therefore, given by

$$\mathrm{R} = \frac{1}{|\mathbb{S}|} \sum_{\mathcal{A}_i \in \mathbb{S}} \gamma_{\mathcal{A}_i}(\mathbb{D}) \tag{21}$$

while the average significance among the selected features is

$$\mathrm{S} = \frac{\sum_{\mathcal{A}_i \neq \mathcal{A}_j \in \mathbb{S}} \{\sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_i) + \sigma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j)\}}{|\mathbb{S}|(|\mathbb{S}| - 1)} \tag{22}$$

i.e.,

$$\mathrm{S} = \frac{\sum_{\mathcal{A}_i \neq \mathcal{A}_j \in \mathbb{S}} 2\gamma_{\{\mathcal{A}_i,\mathcal{A}_j\}}(\mathbb{D}) - \{\gamma_{\mathcal{A}_i}(\mathbb{D}) + \gamma_{\mathcal{A}_j}(\mathbb{D})\}}{|\mathbb{S}|(|\mathbb{S}| - 1)}. \tag{23}$$

Therefore, the problem of selecting a set $\mathbb{S}$ of d relevant and significant features from the whole set $\mathbb{C}$ of m features is equivalent to optimize R and S simultaneously:

$$\max \ \Phi(\mathrm{R}, \mathrm{S}), \quad \Phi = \omega \mathrm{R} + (1 - \omega)\mathrm{S}. \tag{24}$$

where $0 \leq \omega \leq 1$ is a weight parameter and the operator $\Phi(\mathrm{R}, \mathrm{S})$ is defined to combine R and S.

3.2. Computation of relevance and significance

Both relevance and significance of a feature are calculated based on fuzzy-rough set theory. 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 (FEPM) [20], which is denoted by

$$\mathbb{M}_{\mathbb{C}} = \begin{pmatrix} m_{11}^{\mathbb{C}} & m_{12}^{\mathbb{C}} & \cdots & m_{1n}^{\mathbb{C}} \\ m_{21}^{\mathbb{C}} & m_{22}^{\mathbb{C}} & \cdots & m_{2n}^{\mathbb{C}} \\ \cdots & \cdots & \cdots & \cdots \\ m_{c1}^{\mathbb{C}} & m_{c2}^{\mathbb{C}} & \cdots & m_{cn}^{\mathbb{C}} \end{pmatrix}, \tag{25}$$

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.

Definition 1. The relevance of the condition attribute $\mathcal{A}_i$ with respect to the decision attribute set $\mathbb{D}$ can be defined as follows:

$$\gamma_{\mathcal{A}_i}(\mathbb{D}) = \frac{1}{n} \sum_{j=1}^{n} \kappa_j; \quad 0 \leq \gamma_{\mathcal{A}_i}(\mathbb{D}) \leq 1; \quad \text{and}$$

$$\kappa_j = \sup_k \{\sup_s \{\min\{m_{sj}^{\mathcal{A}_i}, \inf_l \{\max\{1 - m_{sl}^{\mathcal{A}_i}, m_{kl}^{\mathbb{D}}\}\}\}\}\}. \tag{26}$$

The family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes. In general, the π function in the one dimensional form is used to assign membership values to different fuzzy equivalence classes for the input features. A fuzzy set with membership function $\pi(x; \bar{c}, \sigma)$ represents a set of points clustered around $\bar{c}$, where

$$\pi(x; \bar{c}, \sigma) = \begin{cases} 2\left(1 - \dfrac{\|x - \bar{c}\|}{\sigma}\right)^2 & \text{for } \dfrac{\sigma}{2} \leq \|x - \bar{c}\| \leq \sigma \\ 1 - 2\left(\dfrac{\|x - \bar{c}\|}{\sigma}\right)^2 & \text{for } 0 \leq \|x - \bar{c}\| \leq \dfrac{\sigma}{2} \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

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, 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 the object x is 0.5 and this is called a crossover point [40].

The $c \times n$ FEPM $\mathbb{M}_{\mathcal{A}_i}$, 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

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

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

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

$$s = \arg \max_j \{m_{kj}^{\mathcal{A}_i}\}, \text{ then } \max_j \{m_{kj}^{\mathcal{A}_i}\} = \max_l \{m_{ls}^{\mathcal{A}_i}\} > 0.$$

In the present work, three fuzzy equivalence classes, namely, low, medium, and high, are considered. Corresponding to three fuzzy sets ($c = 3$), the following relations hold:

$$\bar{c}_1 = \bar{c}_{\text{low}}(\mathcal{A}_i); \quad \bar{c}_2 = \bar{c}_{\text{medium}}(\mathcal{A}_i); \quad \bar{c}_3 = \bar{c}_{\text{high}}(\mathcal{A}_i);$$

$$\sigma_1 = \sigma_{\text{low}}(\mathcal{A}_i); \quad \sigma_2 = \sigma_{\text{medium}}(\mathcal{A}_i); \quad \sigma_3 = \sigma_{\text{high}}(\mathcal{A}_i).$$

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 [41]. Let $\overline{m}_i$ be the mean of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$ along the ith feature $\mathcal{A}_i$. Then $\overline{m}_{i_l}$ and $\overline{m}_{i_h}$ are defined as the means, along the ith feature, of the objects having co-ordinate values in the range $[\mathcal{A}_{i_{\min}}, \overline{m}_i)$ and $(\overline{m}_i, \mathcal{A}_{i_{\max}}]$, 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 [41]:

$$\bar{c}_{\text{low}}(\mathcal{A}_i) = \overline{m}_{i_l}; \quad \bar{c}_{\text{medium}}(\mathcal{A}_i) = \overline{m}_i; \quad \bar{c}_{\text{high}}(\mathcal{A}_i) = \overline{m}_{i_h} \tag{29}$$

$$\begin{aligned} \sigma_{\text{low}}(\mathcal{A}_i) &= 2(\bar{c}_{\text{medium}}(\mathcal{A}_i) - \bar{c}_{\text{low}}(\mathcal{A}_i)) \\ \sigma_{\text{high}}(\mathcal{A}_i) &= 2(\bar{c}_{\text{high}}(\mathcal{A}_i) - \bar{c}_{\text{medium}}(\mathcal{A}_i)) \end{aligned} \tag{30}$$

$$\begin{aligned} \sigma_{\text{medium}}(\mathcal{A}_i) = \frac{\eta}{\mathcal{A}_{i_{\max}} - \mathcal{A}_{i_{\min}}} &[\sigma_{\text{low}}(\mathcal{A}_i)(\mathcal{A}_{i_{\max}} - c_{\text{medium}}(\mathcal{A}_i)) \\ &+ \sigma_{\text{high}}(\mathcal{A}_i)(c_{\text{medium}}(\mathcal{A}_i) - \mathcal{A}_{i_{\min}})] \end{aligned} \tag{31}$$

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 three fuzzy sets. Also, the amount of overlap between three fuzzy sets can be different along the different axis, depending on the distribution of the objects.

To calculate the significance of a condition attribute, the joint relevance $\gamma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D})$ between two attributes $\mathcal{A}_i$ and $\mathcal{A}_j$ needs to be computed. The construction of resultant FEPM $\mathbb{M}_{\{\mathcal{A}_i, \mathcal{A}_j\}}$ is necessary for computing the joint relevance. Let c_i and c_j be the number of fuzzy equivalence classes generated by the condition attributes $\mathcal{A}_i$ and $\mathcal{A}_j$, respectively. If r is the number of resultant fuzzy equivalence partitions, then the $r \times n$ FEPM $\mathbb{M}_{\{\mathcal{A}_i, \mathcal{A}_j\}}$ can be computed as follows:

$$\mathbb{M}_{\{\mathcal{A}_i, \mathcal{A}_j\}} = \mathbb{M}_{\mathcal{A}_i} \cap \mathbb{M}_{\mathcal{A}_j}, \tag{32}$$

where $m_{kt}^{\{\mathcal{A}_i, \mathcal{A}_j\}} = m_{pt}^{\mathcal{A}_i} \cap m_{qt}^{\mathcal{A}_j}$, $k = (p-1)c_j + q$, and $\max\{c_i, c_j\} \le r \le c_i c_j$. In the present work, three fuzzy equivalence classes are considered, that is, $c_i = c_j = 3$.

3.3. Fuzzy-rough MRMS algorithm

Following greedy algorithm is used to solve (24) based on the theory of fuzzy-rough sets:

1. Initialize $\mathbb{C} \leftarrow \{\mathcal{A}_1, \ldots, \mathcal{A}_i, \ldots, \mathcal{A}_j, \ldots, \mathcal{A}_m\}$, $\mathbb{S} \leftarrow \emptyset$.
2. Calculate the centers and radii of three π fuzzy sets for each feature $\mathcal{A}_i \in \mathbb{C}$ according to (29)–(31).
3. Construct the FEPM $\mathbb{M}_{\mathcal{A}_i}$ for each feature $\mathcal{A}_i \in \mathbb{C}$ according to (28).
4. Calculate the relevance $\gamma_{\mathcal{A}_i}(\mathbb{D})$ of each feature $\mathcal{A}_i \in \mathbb{C}$ according to (26).
5. Select feature $\mathcal{A}_i$ as most relevant feature that has highest relevance value $\gamma_{\mathcal{A}_i}(\mathbb{D})$. In effect, $\mathcal{A}_i \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_i$.
6. Repeat the following four steps until $\mathbb{C} = \emptyset$ or the desired number of features d is selected.
7. Construct resultant FEPM $\mathbb{M}_{\{\mathcal{A}_i, \mathcal{A}_j\}}$ for each remaining feature $\mathcal{A}_j \in \mathbb{C}$ and selected feature $\mathcal{A}_i \in \mathbb{S}$ using (32).
8. Calculate the significance of $\mathcal{A}_j \in \mathbb{C}$ with respect to each of the selected features $\mathcal{A}_i \in \mathbb{S}$ as follows:
$$\sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j) = \gamma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}) - \gamma_{\mathcal{A}_i}(\mathbb{D}). \tag{33}$$
9. Remove $\mathcal{A}_j$ from $\mathbb{C}$ if $\sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j) = 0$ for any feature $\mathcal{A}_i \in \mathbb{S}$.
10. From the remaining features of $\mathbb{C}$, select feature $\mathcal{A}_j$ that maximizes the following condition:
$$\omega \gamma_{\mathcal{A}_j}(\mathbb{D}) + \frac{1 - \omega}{|\mathbb{S}|} \sum_{\mathcal{A}_i \in \mathbb{S}} \sigma_{\{\mathcal{A}_i, \mathcal{A}_j\}}(\mathbb{D}, \mathcal{A}_j). \tag{34}$$
As a result of that, $\mathcal{A}_j \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathcal{A}_j$.
11. Stop.

3.4. Computational complexity

The fuzzy-rough set based feature selection method has low computational complexity with respect to the number of features and samples in the original data set. The steps 2, 3, and 4 of the proposed algorithm are executed m times for m attributes. The complexity to compute the centers and radii of three fuzzy sets for each attribute, which is carried out in step 2, is $\mathcal{O}(n)$. The construction of the FEPM of each feature performed in step 3 has $\mathcal{O}(nc)$ time complexity. The computation of the relevance of each feature is carried out in step 4, which has $\mathcal{O}(nc\bar{c})$ time complexity, where $\bar{c}$ represents the number of fuzzy equivalence classes of decision attribute. Hence, the overall time complexity of steps 2, 3, and 4 for m features is $\mathcal{O}(mnc\bar{c})$.

The selection of most relevant feature from the set of m features, which is carried out in step 5, has a complexity $\mathcal{O}(m)$. There is only one loop in step 6 of the proposed feature selection method, which is executed $(d-1)$ times, where d represents the number of selected features. The construction of the resultant FEPM, which is carried out in step 7, and the computation of significance of a candidate feature with respect to an already-selected feature, which is carried out in step 8, have $\mathcal{O}(nc^2)$ and $\mathcal{O}(nc^2\bar{c})$ time complexity, respectively. If $\acute{m} < m$ represents the cardinality of the already-selected feature set, the total complexity of steps 7 and 8 is $\mathcal{O}((m - \acute{m})(nc^2\bar{c}))$. The selection of a feature from $(m - \acute{m})$ candidate features by maximizing both relevance and significance, which is carried out in step 10, has a complexity $\mathcal{O}(m - \acute{m})$. Hence, the total complexity to execute the loop $(d-1)$ times is $\mathcal{O}((d-1)(m - \acute{m})(nc^2\bar{c}))$.

In effect, the selection of a set of d relevant and significant features from the whole set of m features using the proposed

Table 1
Comparative performance analysis of different rough sets on Satimage, Segmentation, and Leukemia II.

Different criteria	Different rough sets	Satimage			Segmentation			Leukemia II		
		K-NN	SVM	C4.5	K-NN	SVM	C4.5	K-NN	SVM	C4.5
Max-Relevance	Classical	69.2/8	67.3/8	67.5/8	63.2/11	57.9/13	63.2/8	81.3/12	81.3/13	81.3/14
	Neighborhood	77.7/13	74.7/10	79.6/17	78.9/13	73.7/11	73.7/12	83.0/25	80.4/11	80.4/21
	Fuzzy	77.9/10	76.4/9	80.0/11	78.9/13	78.9/13	78.9/14	83.0/16	82.1/14	82.1/15
Max-Dependency	Classical	70.4/8	67.3/9	67.5/8	63.2/10	57.9/6	68.4/8	82.1/15	82.1/12	82.1/12
	Neighborhood	83.2/11	81.7/10	81.5/21	79.0/13	73.7/12	73.7/9	85.7/21	83.0/12	83.0/14
	Fuzzy	83.2/12	83.0/11	82.8/15	78.9/13	78.9/13	78.9/13	85.7/17	84.8/13	84.8/13
MRMS	Classical	74.0/9	73.9/10	74.1/10	72.7/9	74.1/13	74.7/12	84.8/15	85.7/15	85.7/15
	Neighborhood	83.4/10	85.0/16	85.2/11	80.5/14	82.6/15	83.2/13	87.5/15	89.3/16	88.4/16
	Fuzzy	84.1/10	84.1/13	84.1/12	80.8/14	84.0/14	85.1/15	88.4/16	90.2/17	89.3/16

fuzzy-rough set based first order incremental search method has an overall computational complexity of $\mathcal{O}(mndc^2\tilde{c})$.

4. Experimental results and discussion

The performance of fuzzy-rough attribute selection method based on the MRMS criterion (fuzzy-rough MRMS) is extensively studied and compared with that of different feature selection and extraction algorithms. The algorithms compared are mutual information based mRMR framework (classical mRMR) [11] and InfoGain [42]; rough set based quick reduct [35] and MRMS framework [17]; fuzzy-rough set based quick reduct [36] and mRMR method (fuzzy-rough mRMR) [20]; margin based approaches such as relevance in estimating features (RELIEF) [43] and iterative search margin based algorithm (SIMBA) [44]; and existing feature extraction algorithms, namely, PCA, ICA, and linear discriminant analysis (LDA) [2]. The performance of fuzzy-rough sets and the MRMS criterion is also compared with that of other feature evaluation criteria, namely, Max-Relevance and Max-Dependency, several existing feature evaluation indices, namely, class separability index [1], DB index [22], Dunn index [23], and FFEI [24], and various rough set models such as classical and neighborhood rough sets.

4.1. Experimental setup

All the algorithms are implemented in C language and run in Ubuntu 11.04 environment with 64 bit support having machine configuration Pentium Core 2 Quad, 2.66 GHz, 4 MB L2 cache, and 4 GB DDR2 RAM. The value of multiplicative parameter η in (31) of fuzzy-rough sets is set to 1.5, while the weight parameter ω in (34) and that of mRMR method are set to 0.5. The discretization method reported in [45] is used to generate equivalence classes of Pawlak's or classical rough sets.

Three pattern classifiers, namely, support vector machine (SVM) [46], K-nearest neighbor (K-NN) rule [2], and C4.5 decision tree [47], are used to evaluate the performance of different dimensionality reduction methods. In the present work, linear kernels are used in the SVM to construct nonlinear decision boundary, while the value of K, chosen for the K-NN rule, is the square root of number of samples in training set. For the data set with small number of features, 80% of total features is selected, while fifty top ranked features is

Table 2
Comparative performance analysis of different rough sets on Colon, Breast I, and Lung.

Different classifiers	Different criteria	Different rough sets	Colon				Breast I				Lung			
			Mean	StDv	Comp	d	Mean	StDv	Comp	d	Mean	StDv	Comp	d
K-NN	Max-Relevance	Classical	58.00	12.59	3.45	2	59.50	8.32	1.24	42	73.20	7.10	2.91	18
		Neighborhood	59.50	11.36	3.36	6	63.50	12.75	0.72	10	76.75	6.50	1.91	13
		Fuzzy	61.19	21.71	1.79	3	66.00	25.03	0.35	3	78.39	7.67	1.30	15
	Max-Dependency	Classical	59.53	24.54	1.81	4	59.50	8.32	1.24	3	77.31	8.45	1.53	8
		Neighborhood	69.52	14.75	0.90	2	61.50	10.34	0.98	11	79.47	8.94	0.90	4
		Fuzzy	68.10	11.18	1.42	2	63.00	10.34	0.81	11	79.47	8.45	0.93	6
	MRMS	Classical	59.50	8.32	3.72	38	59.50	8.32	1.24	3	80.61	7.10	0.67	13
		Neighborhood	63.09	12.59	2.37	7	62.00	8.32	0.95	13	80.61	8.42	0.62	11
		Fuzzy	74.29	8.00		3	70.00	25.39		44	82.84	7.67		43
SVM	Max-Relevance	Classical	64.52	2.62	1.97	10	57.00	10.34	1.30	8	74.50	7.10	2.52	12
		Neighborhood	65.00	9.37	1.59	11	57.50	12.75	1.08	10	76.17	9.25	1.76	13
		Fuzzy	67.61	24.93	0.66	42	59.00	18.86	0.64	4	77.84	7.67	1.46	18
	Max-Dependency	Classical	61.42	14.84	1.87	1	57.50	8.32	1.30	4	77.31	10.21	1.37	4
		Neighborhood	63.09	16.35	1.53	2	56.50	12.75	1.26	3	78.39	7.67	1.30	5
		Fuzzy	64.76	11.10	1.54	1	60.50	8.32	0.65	4	79.47	7.37	1.00	3
	MRMS	Classical	67.61	2.01	1.31	38	57.50	10.34	1.20	3	79.53	7.33	0.98	17
		Neighborhood	70.89	8.29	0.51	10	60.50	10.34	0.60	13	80.61	7.46	0.66	21
		Fuzzy	73.57	14.26		3	63.50	12.03		8	82.83	7.67		2
C4.5	Max-Relevance	Classical	63.09	21.27	1.35	11	61.50	17.51	0.91	14	75.06	9.63	2.53	11
		Neighborhood	64.52	15.89	1.42	8	60.00	15.81	1.14	13	76.75	7.98	2.33	18
		Fuzzy	66.19	11.65	1.38	15	62.50	17.51	0.79	15	78.39	10.21	1.51	21
	Max-Dependency	Classical	61.19	16.03	1.94	8	60.00	8.32	1.35	4	75.06	8.75	2.71	6
		Neighborhood	62.85	15.60	1.70	4	61.00	12.75	1.09	5	77.87	8.45	1.90	3
		Fuzzy	63.09	21.27	1.35	5	63.50	15.81	0.70	3	78.39	8.43	1.73	4
	MRMS	Classical	66.19	11.65	1.38	9	63.50	10.34	0.79	11	79.47	8.43	1.40	14
		Neighborhood	70.89	14.90	0.44	13	66.00	17.51	0.36	8	80.61	7.67	1.11	17
		Fuzzy	73.57	12.31		12	69.00	19.33		14	83.95	5.59		18

Table 3
Comparative performance analysis of different rough sets on Leukemia I, Isolet, and Multiple Features.

Different classifiers	Different criteria	Different rough sets	Leukemia I				Isolet				Multiple Features			
			Mean	StDv	Comp	d	Mean	StDv	Comp	d	Mean	StDv	Comp	d
K-NN	Max-Relevance	Classical	70.72	5.00	1.68	2	58.12	13.28	5.23	17	74.90	8.38	5.24	15
		Neighborhood	71.96	16.13	0.98	8	73.43	8.94	2.58	23	75.65	9.01	4.67	18
		Fuzzy	73.21	14.58	0.84	14	78.39	6.22	1.39	27	80.60	10.42	2.66	10
	Max-Dependency	Classical	69.29	17.45	1.32	5	58.12	14.92	4.74	9	76.10	8.03	5.00	7
		Neighborhood	70.36	16.13	1.22	7	78.94	11.94	0.76	6	74.60	12.73	3.66	4
		Fuzzy	71.96	16.92	0.95	6	79.71	6.38	0.88	4	80.40	9.22	3.04	6
	MRMS	Classical	71.96	19.85	0.86	7	67.09	13.22	3.29	17	86.10	8.31	1.31	17
		Neighborhood	74.64	16.13	0.58	12	79.17	11.32	0.73	13	87.40	9.03	0.79	12
		Fuzzy	78.57	13.88		7	82.12	5.83		44	89.80	3.26		31
SVM	Max-Relevance	Classical	65.00	23.63	0.55	8	58.07	14.21	5.20	21	81.10	8.93	3.65	12
		Neighborhood	66.43	18.38	0.46	10	74.17	11.92	2.06	18	80.20	10.25	3.49	9
		Fuzzy	66.61	16.71	0.46	9	78.34	7.31	1.53	12	82.80	9.21	3.00	17
	Max-Dependency	Classical	67.14	16.13	0.39	10	58.07	13.63	5.41	5	79.60	8.03	4.55	16
		Neighborhood	65.00	14.20	0.72	12	78.53	7.21	1.47	9	80.60	11.27	3.09	8
		Fuzzy	63.57	17.24	0.85	45	79.49	4.29	1.54	11	83.45	12.03	2.19	4
	MRMS	Classical	67.14	16.13	0.39	12	74.17	5.94	3.62	22	89.75	4.38	1.35	17
		Neighborhood	67.68	15.53	0.32	14	79.29	4.33	1.65	21	90.65	4.26	0.85	13
		Fuzzy	70.00	16.60		48	82.35	3.99		41	92.10	3.34		28
C4.5	Max-Relevance	Classical	69.29	17.45	0.89	9	57.98	16.25	2.72	8	76.10	6.93	3.85	15
		Neighborhood	70.36	17.17	0.75	14	64.65	12.66	1.94	8	78.20	10.73	2.21	19
		Fuzzy	71.96	16.13	0.56	11	65.06	11.04	2.00	7	79.55	9.22	2.09	9
	Max-Dependency	Classical	68.93	17.96	0.92	5	57.98	16.25	2.72	8	76.45	8.93	3.10	13
		Neighborhood	69.29	16.13	0.92	2	64.65	12.66	1.94	8	84.40	10.71	0.55	9
		Fuzzy	71.96	19.85	0.50	7	65.06	11.04	2.00	7	79.35	9.04	2.18	5
	MRMS	Classical	70.36	15.35	0.79	12	66.51	9.28	1.84	17	85.30	11.21	0.30	18
		Neighborhood	73.39	22.91	0.30	15	72.03	11.13	0.64	14	87.30	8.77	−0.27	8
		Fuzzy	76.07	16.82		16	75.32	11.94		38	86.45	4.93		19

considered for the data set with large number of features. In all cases, the result is presented for highest classification accuracy.

4.2. Description of data sets

This section reports some benchmark data sets that are used to evaluate the performance of different methods. While Satimage, Segmentation, Isolet, and Multiple Features data sets are downloaded from the *UCI Machine Learning Repository* [48], Breast Cancer I, Leukemia I, Colon Cancer, Lung Cancer, and Leukemia II data sets are available at the *Kent Ridge Bio-medical Data Set Repository* [49].

1. *Satimage:* The database is a tiny sub-area of a scene, consisting of 82×100 pixels, each pixel covering an area on the ground of approximately 80×80 m. The information given for each pixel consists of the class value and the intensities in four spectral bands, from the green, red, and infra-red regions of the spectrum. The data set contains 6435 examples: 4435 training and 2000 testing, with 36 real valued attributes and 6 classes.
2. *Breast Cancer I:* This data set contains expression levels of 7129 genes in 49 breast tumor samples. The samples are classified according to their estrogen receptor (ER) status: 25 samples are ER positive while other 24 samples are ER negative.
3. *Leukemia I:* 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.
4. *Colon Cancer:* The colon cancer data set contains expression levels of 2000 genes and 62 samples from two classes: 40 tumor and 22 normal colon tissues.
5. *Lung Cancer:* This data set contains 181 tissue samples: among them 31 are malignant pleural mesothelioma and rest 150 adenocarcinoma of the lung. Each sample is described by the expression levels of 12,533 genes.

Table 4
Comparative performance analysis of different indices on Satimage, Segmentation, and Leukemia II.

Different criteria	Different indices	Satimage			Segmentation			Leukemia II		
		K-NN	SVM	C4.5	K-NN	SVM	C4.5	K-NN	SVM	C4.5
Max-Relevance	Class separability	60.3/7	57.2/6	59.2/7	55.5/6	46.1/5	58.2/7	75.0/11	76.8/12	77.7/13
	DB index	77.2/10	75.7/11	77.1/10	72.0/9	76.0/11	77.8/10	77.7/11	79.5/13	79.5/14
	Dunn index	56.0/7	49.2/6	50.6/7	55.5/7	47.7/7	57.8/8	80.4/13	81.3/14	79.5/14
	FFEI	60.3/8	54.3/6	58.6/7	77.3/10	81.3/14	82.8/13	71.4/10	77.7/12	75.0/13
Max-Dependency	Class separability	60.3/7	57.2/7	58.6/6	52.6/5	46.1/5	57.9/6	75.0/9	75.0/12	76.8/11
	DB index	79.5/10	75.3/9	78.9/11	68.4/8	73.7/10	78.9/11	78.6/10	75.0/10	79.5/11
	Dunn index	56.0/7	51.8/8	50.6/6	55.5/7	47.4/6	57.9/7	71.4/9	77.7/12	78.6/10
	FFEI	60.3/7	54.3/7	56.4/7	57.9/7	73.7/11	78.9/10	80.4/13	75.9/11	74.1/10
MRMS	Class separability	60.3/7	57.2/7	59.2/7	55.5/7	46.1/5	58.2/7	75.0/11	76.8/11	77.7/12
	DB index	83.3/10	81.2/9	83.4/11	72.0/9	76.0/11	77.8/11	80.4/12	81.3/13	79.5/12
	Dunn index	56.0/8	49.2/6	50.6/7	55.5/9	47.7/6	57.8/9	80.4/12	81.3/13	79.5/13
	FFEI	60.3/7	54.3/7	58.6/8	78.8/13	83.0/15	82.8/14	71.4/9	77.7/10	75.0/12
	Fuzzy-rough	84.1/10	84.1/13	84.1/12	80.8/14	84.0/14	85.1/15	88.4/16	90.2/17	89.3/16

Table 5
Comparative performance analysis of different indices on Colon, Breast I, and Lung.

Different classifiers	Different criteria	Different indices	Colon				Breast I				Lung			
			Mean	StDv	Comp	d	Mean	StDv	Comp	d	Mean	StDv	Comp	d
K-NN	Max-Relevance	Class separability	68.24	13.29	1.23	12	58.50	8.32	1.36	11	75.06	8.85	2.10	10
		DB index	70.95	14.67	0.63	2	59.00	11.65	1.25	12	76.75	7.10	1.84	13
		Dunn index	71.18	4.53	1.07	4	59.50	13.01	1.16	7	76.75	8.01	1.73	16
		FFEI	69.62	11.24	1.07	13	59.00	11.65	1.48	10	75.06	7.56	2.28	14
	Max-Dependency	Class separability	65.21	14.22	1.76	3	57.50	8.32	1.33	4	75.06	7.55	2.29	5
		DB index	66.83	11.25	1.71	2	59.00	12.48	1.23	2	75.06	7.55	2.29	4
		Dunn index	66.83	11.25	1.71	4	58.50	10.34	1.33	3	77.87	6.50	1.56	6
		FFEI	65.38	10.66	2.11	3	58.00	12.48	1.34	3	75.06	7.13	2.35	10
	MRMS	Class separability	72.62	10.95	0.39	5	57.50	10.34	1.44	10	76.75	7.83	1.75	11
		DB index	72.38	13.84	0.38	3	61.50	10.34	0.98	16	76.75	5.59	2.03	10
		Dunn index	72.38	13.84	0.38	4	61.50	12.48	0.95	13	77.87	10.21	1.23	14
		FFEI	70.95	11.46	0.75	7	59.00	13.01	1.22	18	76.17	7.67	1.95	18
		Fuzzy-rough	74.29	8.00		3	70.00	25.39		44	82.84	7.67		43
SVM	Max-Relevance	Class separability	68.81	14.61	0.74	8	57.00	12.75	1.17	13	74.50	10.55	2.02	15
		DB index	70.71	15.51	0.43	13	57.50	12.75	1.08	17	76.17	9.80	1.70	17
		Dunn index	70.71	15.51	0.43	14	57.50	18.86	0.85	12	76.72	8.85	1.65	19
		FFEI	69.05	14.89	0.69	17	57.00	12.75	1.17	16	75.06	7.67	2.27	10
	Max-Dependency	Class separability	67.38	16.07	0.91	18	56.50	12.48	1.28	3	75.06	8.43	7.00	5
		DB index	68.81	14.61	0.74	12	56.00	13.01	1.34	4	76.17	10.55	1.62	6
		Dunn index	69.05	14.89	0.69	9	57.00	12.75	1.17	2	77.31	9.80	1.40	3
		FFEI	67.38	13.84	0.98	10	56.00	8.32	1.62	5	75.06	8.43	2.16	6
	MRMS	Class separability	72.14	11.12	0.25	10	56.50	8.32	1.51	10	75.61	8.85	1.95	14
		DB index	72.38	13.84	0.19	4	58.00	12.48	1.00	17	76.75	7.67	1.77	15
		Dunn index	72.38	13.84	0.19	4	57.50	8.32	1.30	5	77.87	7.60	1.46	21
		FFEI	71.90	13.96	0.26	6	57.00	13.01	1.16	6	76.17	8.43	1.85	13
		Fuzzy-rough	73.57	14.26		3	63.50	12.03		8	82.83	7.67		2
C4.5	Max-Relevance	Class separability	69.76	12.14	0.70	14	60.50	12.75	1.16	13	75.06	8.01	2.88	13
		DB index	70.95	12.89	0.46	13	63.50	8.32	0.83	11	76.17	7.56	2.62	11
		Dunn index	70.95	12.14	0.48	9	63.50	11.65	0.77	12	77.87	6.50	2.24	20
		FFEI	69.05	11.40	0.85	3	61.00	8.32	1.20	7	75.06	7.13	3.10	17
	Max-Dependency	Class separability	66.19	14.05	1.25	4	59.50	12.48	1.31	9	75.61	8.01	2.70	6
		DB index	67.38	10.83	1.19	3	58.00	13.01	1.49	4	77.87	7.56	2.05	8
		Dunn index	68.09	11.18	1.04	5	58.50	12.75	1.43	5	77.87	8.42	1.90	4
		FFEI	66.19	12.59	1.33	3	59.50	8.32	1.43	3	75.06	7.97	2.89	6
	MRMS	Class separability	71.90	14.05	0.28	10	63.00	8.32	0.90	10	77.87	7.83	2.00	14
		DB index	72.38	15.59	0.19	13	65.00	12.48	0.55	13	78.42	5.59	2.21	18
		Dunn index	72.38	13.22	0.21	5	64.50	12.48	0.62	14	78.42	8.36	1.74	24
		FFEI	71.90	14.05	0.28	13	63.50	10.34	0.79	11	76.17	8.85	2.35	18
		Fuzzy-rough	73.57	12.31		12	69.00	19.33		14	83.95	5.59		18

6. *Leukemia II:* This data set consists of gene expression profiles of 215 training and 112 testing samples classified into 7 classes, six subtypes of pediatric acute lymphoblastic leukemia and one that contains diagnostic samples that did not fit into any one of the six groups. The data set contains total 12,558 genes.
7. *Isolet:* The data set consists of several spectral coefficients of utterances of English alphabets by 150 subjects. There are 617 real valued features with 7797 instances and 26 classes.
8. *Multiple Features:* Multiple features data set consists of features of handwritten numerals ('0'-'9') extracted from a collection of Dutch utility maps. 200 patterns per class (for a total of 2,000 patterns) have been digitized in binary images. Total 649 attributes are there in the data set.
9. *Segmentation:* This data set contains instances that are drawn randomly from a database of 7 outdoor images. The images are hand segmented to create a classification for every pixel, where each instance is a 3×3 region. The data set contains 3310 examples: 210 training and 2100 testing, with 18 continuous attributes and 7 classes.

To compute the classification accuracy of the C4.5, K-NN, and SVM, both training-testing and 10 fold cross-validation (CV) are performed. The 10-fold CV is performed on Breast I, Colon, Lung, Leukemia I, Isolet, and Multiple Features data sets, while the training-testing is done on Satimage, Leukemia II, and Segmentation data sets.

4.3. Statistical significance test

In case of 10-fold CV, the means and standard deviations of the classification accuracy of the SVM, C4.5, and K-NN are computed for Breast I, Colon, Lung, Leukemia I, Isolet, and Multiple Features data sets. Tests of significance are performed for the inequality of means (of the classification accuracy of the SVM, C4.5, and K-NN) obtained using the fuzzy-rough MRMS method and other approaches. Since both mean pairs and variance pairs are unknown and different, a generalized version of t-test is used here. The above problem is the classical Behrens–Fisher problem in hypothesis testing. The test statistic, described and tabled in [50], is of the form

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\lambda_1 {\sigma_1}^2 + \lambda_2 {\sigma_2}^2}}, \tag{35}$$

where μ_1, μ_2 are the means, σ_1, σ_2 the standard deviations, and $\lambda_1 = 1/n_1$, $\lambda_2 = 1/n_2$, n_1, n_2 are number of observations. Tables 2, 3, 5, 6, 8 and 9 report the individual means and standard deviations, and the value of 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.

Table 6
Comparative performance analysis of different indices on Leukemia I, Isolet, and Multiple Features.

Different classifiers	Different criteria	Different indices	Leukemia I				Isolet				Multiple Features			
			Mean	StDv	Comp	d	Mean	StDv	Comp	d	Mean	StDv	Comp	d
K-NN	Max-Relevance	Class separability	67.50	18.30	1.52	12	45.33	16.93	6.50	11	72.60	10.71	4.86	11
		DB index	68.93	20.33	1.24	10	73.26	12.33	2.06	14	73.90	8.31	5.63	18
		Dunn index	68.93	14.20	1.53	15	73.11	11.57	2.20	12	73.90	9.03	5.24	9
		FFEI	67.68	15.53	1.65	13	63.88	10.34	4.86	16	73.20	9.26	5.35	14
	Max-Dependency	Class separability	66.61	16.71	1.74	9	45.09	16.22	6.79	7	71.90	9.26	5.77	5
		DB index	67.86	19.07	1.44	7	68.17	12.64	3.17	6	73.90	9.03	5.24	8
		Dunn index	69.29	17.45	1.32	8	69.42	12.43	2.93	6	72.70	8.38	6.01	4
		FFEI	67.86	17.26	1.53	6	63.88	10.93	4.66	8	72.40	9.22	5.63	6
	MRMS	Class separability	71.96	19.85	0.86	13	45.20	17.53	6.32	15	80.40	9.03	3.10	17
		DB index	73.21	20.22	0.69	10	73.11	10.38	2.40	21	82.05	10.71	2.19	13
		Dunn index	74.64	20.22	0.51	12	72.89	9.37	2.65	16	82.10	8.38	2.71	23
		FFEI	70.36	16.13	1.22	14	63.88	11.26	4.55	19	80.60	8.31	3.26	18
		Fuzzy-rough	78.57	13.88		7	82.12	5.83		44	89.80	3.26		31
SVM	Max-Relevance	Class separability	60.71	18.13	1.19	13	45.25	15.65	7.27	13	76.90	12.27	3.78	12
		DB index	63.75	21.59	0.72	16	73.23	11.52	2.37	16	80.10	9.22	3.87	14
		Dunn index	63.75	16.13	0.85	10	72.62	10.25	2.80	19	79.95	10.71	3.42	21
		FFEI	61.96	16.95	1.07	14	71.01	11.77	2.89	10	78.95	8.31	4.64	19
	Max-Dependency	Class separability	62.68	16.95	0.98	6	45.25	16.21	7.03	6	78.40	8.38	4.80	6
		DB index	64.82	14.36	0.74	4	69.74	9.32	3.93	8	80.30	7.22	4.69	3
		Dunn index	64.82	17.29	0.68	3	69.80	8.31	4.31	4	80.20	9.03	3.91	8
		FFEI	62.68	14.36		4	71.01	8.33	3.88	6	79.15	9.01	4.26	4
	MRMS	Class separability	64.82	17.29	0.68	13	45.30	18.21	6.29	11	80.70	8.31	4.03	18
		DB index	65.18	15.35	0.67	18	73.37	8.21	3.11	17	85.30	10.71	1.92	9
		Dunn index	65.18	19.98	0.59	11	72.72	7.66	3.53	21	85.35	9.26	2.17	32
		FFEI	63.93	17.53	0.80	14	69.87	8.52	4.19	22	82.50	13.27	2.22	17
		Fuzzy-rough	70.00	16.60		48	82.35	3.99		41	92.10	3.34		28
C4.5	Max-Relevance	Class separability	67.50	16.92	1.14	14	45.43	12.53	5.46	21	74.30	9.01	3.74	16
		DB index	68.93	17.45	0.93	11	62.60	10.33	2.55	13	77.35	7.22	3.29	19
		Dunn index	68.93	19.85	0.87	8	54.93	11.43	3.90	18	77.40	9.03	2.78	11
		FFEI	67.68	17.17	1.11	15	52.01	9.42	4.85	16	76.70	8.38	3.17	10
	Max-Dependency	Class separability	65.00	20.55	1.32	8	45.43	13.28	5.29	7	73.75	11.27	3.26	5
		DB index	67.86	16.13	1.12	6	60.84	14.27	2.46	6	75.80	10.71	2.86	8
		Dunn index	68.93	16.13	0.97	5	54.93	11.84	3.83	18	76.25	9.26	3.07	4
		FFEI	66.61	16.71	1.26	7	52.40	12.53	4.19	6	75.20	8.31	3.68	10
	MRMS	Class separability	70.36	16.13	0.78	9	45.52	16.27	4.67	12	83.30	9.26	0.95	19
		DB index	71.96	19.85	0.50	10	63.92	9.27	2.38	31	87.15	7.22	−0.25	8
		Dunn index	71.96	17.83	0.53	11	54.93	10.22	4.10	18	86.20	8.38	0.08	17
		FFEI	70.36	16.71	0.76	12	52.40	11.23	4.42	6	87.30	9.01	−0.26	21
		Fuzzy-rough	76.07	16.82		16	75.32	11.94		38	86.45	4.93		19

4.4. Performance of various rough set models

In 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. Different feature evaluation criteria such as Max-Dependency, Max-Relevance, and MRMS are considered for feature selection. In this regard, it should be mentioned that the classical or Pawlak's rough set based feature selection method reported in [17] uses the MRMS criterion, while the quick reduct [35], fuzzy-rough quick reduct [36], and neighborhood quick reduct [39] algorithms select features using Max-Dependency criterion.

Tables 1–3 present the comparative performance of different rough set models for attribute selection task. The results and subsequent discussions are presented in these tables with respect to the classification accuracy of the K-NN, SVM, and C4.5. From the results reported in Table 1, it can be seen that the fuzzy-rough MRMS method attains maximum classification accuracy of the K-NN, SVM, and C4.5 in most of the cases. Out of 9 cases of training-testing, the fuzzy-rough MRMS method achieves highest classification accuracy in 7 cases, while neighborhood rough set based MRMS method attains it only in 2 cases.

On the other hand, among the 144 comparisons of 10-fold CV reported in Tables 2 and 3, the MRMS criterion with fuzzy-rough sets provides significantly better results in 47 cases and better results but not significantly in 96 cases, while better result is achieved only in 1 case using neighborhood rough sets based on the MRMS criterion. In brief, out of total 27 cases, the fuzzy-rough sets and neighborhood rough sets attain highest classification accuracy in 24 and 3 cases, respectively, based on the MRMS criterion.

Following conclusions can be drawn from the results reported in Tables 1–3:

- the performance of the MRMS criterion is significantly better than that of other criteria, namely, Max-Dependency and Max-Relevance, irrespective of rough set models used;
- the performance of fuzzy-rough sets is significantly better than that of classical and neighborhood rough sets, irrespective of the feature evaluation criteria used; and
- the MRMS criterion based feature selection method using fuzzy-rough sets achieves higher classification accuracy in most of the cases, irrespective of the data sets, feature evaluation criteria, rough set models, experimental setup, and classifiers used.

The better performance of the fuzzy-rough MRMS method is achieved due to the fact that the MRMS criterion can identify relevant and significant features from high dimensional real life data sets more efficiently than Max-Dependency and Max-Relevance criteria, while the fuzzy-rough sets can capture uncertainties associated with the data more accurately. In effect, a reduced set of relevant and significant features is obtained using fuzzy-rough

Table 7
Comparative performance analysis of different methods on Satimage, Segmentation, and Leukemia II.

Different methods/algorithms	Satimage			Segmentation			Leukemia II		
	K-NN	SVM	C4.5	K-NN	SVM	C4.5	K-NN	SVM	C4.5
InfoGain	71.4/13	72.2/9	71.8/15	72.4/14	72.6/9	72.3/11	82.7/5	82.7/7	83.2/10
Classical mRMR	75.5/9	75.4/10	75.4/10	72.8/10	73.8/10	74.3/10	84.8/11	84.8/10	84.8/11
Fuzzy-rough mRMR	84.0/11	84.6/13	83.7/12	80.3/11	84.1/12	84.7/12	87.5/13	89.3/12	90.2/16
RELIEF	83.3/23	81.7/16	82.2/28	73.7/5	82.4/7	86.3/13	87.5/10	85.7/12	84.8/9
SIMBA	83.2/22	81.5/21	82.7/20	74.8/3	84.3/9	84.3/11	87.5/13	89.3/15	85.7/12
PCA	82.6/8	84.0/8	82.0/9	78.9/7	89.5/9	94.7/14	80.4/12	78.6/10	79.5/12
ICA	83.3/10	83.5/12	82.2/11	75.1/9	90.0/13	90.1/11	85.2/14	85.7/11	84.8/15
LDA	82.7/9	82.7/10	82.4/10	82.0/12	90.3/14	89.6/10	86.6/15	87.4/16	84.8/14
Fuzzy-rough MRMS	84.1/10	84.1/13	84.1/12	80.8/14	84.0/14	85.1/15	88.4/16	90.2/17	89.3/16

MRMS algorithm with significantly lesser time as reported in Table 10.

4.5. Performance of different feature evaluation indices

In order to establish the effectiveness of fuzzy-rough sets over other feature evaluation indices, extensive experimentation is done on different real life data sets. Tables 4–6 present the comparative performance of fuzzy-rough MRMS method and various feature evaluation indices such as class separability index, DB index, Dunn index, and FFEI considering different feature evaluation criteria such as Max-Relevance, Max-Dependency, and MRMS.

From the results reported in Table 4, it can be seen that the fuzzy-rough MRMS method attains highest classification accuracy on Satimage, Leukemia II, and Segmentation data sets, irrespective of the classifiers used. Tables 5 and 6 report the comparative performance in case of 10-fold CV on Colon, Breast I, Lung, Leukemia I, Isolet, and Multiple Features data sets. The results and subsequent discussions are analyzed in these tables with respect to the classification accuracy of the K-NN, SVM, and C4.5. All the results reported in Tables 5 and 6 confirm that the fuzzy-rough MRMS method provides significantly better results in 94 cases and better but not significantly in 120 cases out of total 216 cases. On the other hand, both DB index and FFEI based on the MRMS criterion achieve better results, but not significantly, than that of the proposed fuzzy-rough MRMS method in only 1 case each. The results reported in Tables 4–6 also establish the fact that the performance of the MRMS criterion is better than that of other two criteria, namely, Max-Dependency and Max-Relevance, irrespective of feature evaluation indices, classifiers, and data sets used.

4.6. Performance of different algorithms

Finally, Tables 7–10 compare the performance of the proposed fuzzy-rough MRMS algorithm with that of different existing feature selection and extraction algorithms on various data sets. From the results reported in Table 7, it is seen that the fuzzy-rough MRMS algorithm achieves highest classification accuracy of the SVM, C4.5, and K-NN in 4 cases out of total 9 cases, while the PCA, LDA, and fuzzy-rough mRMR attain highest classification accuracy in only 1, 2, and 2 cases, respectively. Tables 8 and 9 report the performance of different methods in case of 10-fold CV, along with the results of test of significance, for the K-NN, SVM, and C4.5. From the results reported in these tables, it can be seen that the proposed

Table 8
Comparative performance analysis of different methods on Colon, Breast I, and Lung.

Different classifiers	Different methods/algorithms	Colon				Breast I				Lung			
		Mean	StDv	Comp	d	Mean	StDv	Comp	d	Mean	StDv	Comp	d
	InfoGain	67.86	17.82	1.04	2	56.00	27.97	1.17	7	68.10	11.18	3.44	37
	Classical mRMR	67.86	16.12	1.13	46	59.50	18.33	1.06	6	68.10	11.18	3.44	37
	Fuzzy-rough mRMR	71.43	20.26	0.42	5	56.00	20.66	1.35	3	82.84	7.67	0.00	37
	RELIEF	65.71	15.07	1.59	6	59.50	8.32	1.24	13	74.53	7.62	2.43	17
K-NN	SIMBA	66.19	14.89	1.51	14	61.00	9.83	1.05	16	76.75	9.07	1.62	19
	PCA	68.88	15.24	0.99	4	63.50	14.38	0.70	9	79.47	10.21	0.83	11
	ICA	67.68	16.82	1.12	7	63.50	15.93	0.69	12	80.61	8.45	0.62	10
	LDA	67.86	11.35	1.46	10	63.00	13.37	0.77	9	80.06	8.43	0.77	13
	Fuzzy-rough MRMS	74.29	8.00		3	70.00	25.39		44	82.84	7.67		43
	InfoGain	68.33	16.91	0.75	14	60.00	28.28	0.36	12	67.25	34.12	1.41	49
	Classical mRMR	72.38	11.40	0.21	11	58.00	33.27	3.22	9	76.20	8.85	1.79	17
	Fuzzy-rough mRMR	73.09	19.38	0.06	10	62.00	23.94	9.46	26	77.41	8.45	1.50	19
	RELIEF	65.28	12.31	1.39	10	59.50	7.55	0.89	10	75.09	8.07	2.20	13
SVM	SIMBA	67.01	11.23	1.14	13	59.50	8.26	0.87	9	75.64	8.04	2.05	10
	PCA	68.24	10.63	0.95	11	62.00	13.37	0.26	9	77.31	7.70	1.61	10
	ICA	66.97	11.44	1.14	6	61.00	13.70	0.43	7	76.20	7.53	1.95	9
	LDA	67.56	9.89	1.09	9	62.00	9.43	0.31	8	77.41	9.03	1.45	12
	Fuzzy-rough MRMS	73.57	14.26		3	63.50	12.03		8	82.83	7.67		2
	InfoGain	66.19	11.65	1.38	12	57.50	14.29	1.51	13	68.01	11.70	3.89	9
	Classical mRMR	72.38	11.40	0.22	11	59.00	11.32	1.41	12	77.86	8.35	1.91	11
	Fuzzy-rough mRMR	73.09	16.30	0.07	9	57.50	15.94	1.45	14	78.39	8.94	1.66	15
	RELIEF	64.76	18.02	1.28	13	59.00	13.99	1.33	11	75.64	7.37	2.84	12
C4.5	SIMBA	66.19	19.56	1.01	14	60.50	12.03	1.18	12	76.20	8.45	2.42	15
	PCA	69.76	12.14	0.70	11	63.50	22.74	0.58	6	78.98	7.83	1.63	6
	ICA	68.24	13.57	0.92	9	63.00	23.52	0.62	7	77.86	7.46	2.06	7
	LDA	69.76	17.04	0.57	9	61.50	16.47	0.93	12	78.98	7.56	1.67	9
	Fuzzy-rough MRMS	73.57	12.31		12	69.00	19.33		14	83.95	5.59		18

Table 9
Comparative performance analysis of different methods on Leukemia I, Isolet, and Multiple Features.

Different classifiers	Different methods/algorithms	Leukemia I				Isolet				Multiple Features			
		Mean	StDv	Comp	d	Mean	StDv	Comp	d	Mean	StDv	Comp	d
K-NN	InfoGain	72.14	18.83	0.87	20	76.09	11.38	1.49	26	76.70	5.03	6.91	21
	Classical mRMR	70.89	16.79	1.11	18	77.04	9.11	1.49	15	79.75	9.37	3.20	14
	Fuzzy-rough mRMR	73.22	4.53	1.16	10	82.06	6.64	0.03	23	83.75	10.21	1.79	13
	RELIEF	71.96	16.71	0.96	10	78.32	7.10	1.31	16	81.80	8.38	2.81	16
	SIMBA	73.39	17.26	0.74	11	78.38	4.87	1.56	11	82.90	8.27	2.45	13
	PCA	74.64	17.83	0.55	8	76.21	5.38	2.36	10	84.15	7.22	2.26	8
	ICA	74.64	14.00	0.63	12	78.27	6.44	1.40	13	85.10	8.05	1.71	11
	LDA	76.07	16.82	0.36	10	78.43	5.18	1.50	14	84.65	9.21	1.67	9
	Fuzzy-rough MRMS	78.57	13.88		7	82.12	5.83		44	89.80	3.26		31
SVM	InfoGain	60.00	28.28	0.96	12	76.47	13.83	1.29	16	79.50	10.32	3.67	16
	Classical mRMR	65.00	18.58	0.63	16	75.63	12.48	1.62	12	83.10	9.26	2.89	15
	Fuzzy-rough mRMR	68.57	19.98	0.17	49	81.92	8.33	0.15	21	88.60	8.31	1.24	21
	RELIEF	66.43	19.57	0.44	10	79.18	8.19	1.10	14	83.90	10.33	2.39	15
	SIMBA	67.68	14.00	0.34	21	79.43	7.33	1.11	13	85.20	9.51	2.16	13
	PCA	69.29	17.45	0.09	11	80.76	8.82	0.52	9	85.70	8.05	2.32	7
	ICA	70.36	16.13	−0.05	9	83.25	5.03	−0.44	7	84.95	10.27	2.09	11
	LDA	70.71	16.92	−0.10	14	82.07	6.18	0.12	12	85.75	8.21	2.27	9
	Fuzzy-rough MRMS	70.00	16.60		48	82.35	3.99		41	92.10	3.34		28
C4.5	InfoGain	70.36	16.13	0.78	7	71.48	9.38	0.80	14	78.50	9.26	2.40	19
	Classical mRMR	70.54	20.91	0.65	8	71.54	10.42	0.75	19	80.35	10.44	1.67	21
	Fuzzy-rough mRMR	73.21	19.85	0.35	11	73.48	7.44	0.42	22	81.05	9.04	1.66	9
	RELIEF	71.79	20.49	0.51	9	71.75	6.06	0.85	18	80.80	9.26	1.70	14
	SIMBA	71.96	19.85	0.50	10	72.12	7.27	0.73	11	81.30	8.38	1.68	16
	PCA	73.39	22.91	0.30	8	66.88	8.22	1.84	12	85.20	11.28	0.32	11
	ICA	74.64	17.83	0.18	7	80.76	6.74	−1.25	9	87.80	9.22	−0.41	9
	LDA	74.64	14.00	0.21	7	82.25	6.11	−1.63	14	88.90	5.27	−1.07	10
	Fuzzy-rough MRMS	76.07	16.82		16	75.32	11.94		38	86.45	4.93		19

method attains significantly better results than other algorithms in 28 cases out of total 144 cases and better results but not significantly in 109 cases, while better results, but not significantly, are achieved by the ICA and LDA in 4 and 3 cases, respectively. In brief, out of total 27 cases, the fuzzy-rough MRMS method attains highest classification accuracy in 18 cases, while the PCA, ICA, LDA, and fuzzy-rough mRMR methods achieve it only 1, 1, 5, and 2 cases, respectively.

Table 10
Execution time (in second) of different methods for various data sets.

Different criteria	Different indices	Different benchmark data sets								
		Colon	Breast I	Lung	Leukemia I	Isolet	Mult.Feat.	Satimage	Segmentation	Leukemia II
Max-Relevance	Class Separ	0.26	0.38	2.77	0.21	1.26	0.28	0.08	0.04	6.28
	DB index	0.28	0.39	1.28	0.32	1.66	0.48	0.12	0.03	2.10
	Dunn index	0.19	0.22	0.96	0.28	1.72	0.37	0.11	0.03	2.08
	FFEI	2.98E03	3.43E03	9.83E03	3.83E03	1.83E04	4.78E03	2.52E03	0.04	7.25E03
	Classical	0.20	0.31	2.00E01	0.80	8.30	3.30	0.10	0.10	5.21
	Neighborhood	5.40E02	6.30E02	8.10E04	7.30E02	2.10E04	1.00E04	8.40E02	6.30E02	5.70E04
	Fuzzy-rough	0.20	2.21	8.41	2.10	2.20E01	7.41	1.70E01	1.60E01	2.20E01
Max-Dependency	Class Separ	1.77	1.98	3.22	2.04	4.28	2.83	0.45	0.05	7.68
	DB index	2.34	1.82	3.82	2.16	5.02	3.92	0.84	0.04	2.98
	Dunn index	1.92	1.55	3.21	1.93	4.88	4.27	0.78	0.04	3.12
	FFEI	1.19E04	1.43E04	1.38E04	1.27E04	1.87E04	1.53E04	8.34E03	0.06	8.25E03
	Classical	2.60E01	1.40E01	8.30E01	1.80E01	1.30E02	4.60E01	4.91	3.30	9.40E02
	Neighborhood	1.20E04	9.30E03	6.20E05	1.10E04	2.20E05	6.70E04	2.40E04	2.80E04	5.30E05
	Fuzzy-rough	9.80E03	9.80E03	5.90E05	9.40E03	2.60E05	5.20E04	3.30E04	2.70E04	5.50E05
MRMS	Class Separ	0.52	0.62	3.82	0.42	1.52	0.42	0.46	0.25	6.12E02
	DB index	0.37	0.62	2.55	0.41	2.81	0.61	2.09	0.22	1.12E03
	Dunn index	0.29	0.44	1.28	0.45	3.12	0.52	2.62	0.20	6.01E02
	FFEI	4.32E03	4.84E03	1.23E04	5.27E03	2.13E04	6.72E03	5.33E04	4.25E01	7.53E04
	Classical	0.20	0.40	4.30E01	1.52	9.40	5.92	0.31	0.24	7.90
	Neighborhood	7.30E02	8.20E02	9.20E04	9.20E02	3.40E04	2.10E04	1.00E03	6.50E02	8.70E04
InfoGain		0.51	5.39	1.71E01	5.40	7.02	0.36	0.09	0.05	4.72
Classical mRMR		0.52	5.51	1.80E01	5.40	7.40	0.40	0.10	0.10	5.00
Fuzzy-rough mRMR		1.71	6.10E01	3.80E01	3.10E01	5.60E01	1.40E01	1.20	1.80	6.50
RELIEF		2.06	5.28E01	2.17E01	1.83E01	3.26E01	1.24E01	2.77	1.29	4.22
SIMBA		3.39E02	5.92E02	6.83E02	4.59E02	3.02E03	2.01E03	1.02E03	3.82E02	4.11E03
PCA		2.20E04	3.00E04	3.20E04	3.80E04	1.10E04	1.80E04	7.80E03	8.00E02	4.00E03
ICA		6.29E02	7.22E02	8.47E02	5.29E02	4.11E03	3.11E03	1.23E03	4.51E02	5.63E03
LDA		3.18E02	5.11E02	1.13E03	4.18E02	8.31E02	2.18E03	9.81E02	4.92E02	8.92E02
Fuzzy-rough MRMS		0.40	2.81	9.52	2.71	3.20E01	1.00E01	4.30E01	2.30E01	2.70E01

All the results reported in Tables 7–9 also establish the fact that the mRMR criterion based feature selection method in fuzzy approximation spaces (fuzzy-rough mRMR) [20] improves the classification accuracy significantly over its crisp counterpart (classical mRMR) [11], irrespective of the classifiers and data sets used. Out of total 27 cases, the fuzzy-rough mRMR method provides better accuracy than that of classical mRMR in 25 cases. Only for Breast Cancer I data set, the classical mRMR performs better with respect to both C4.5 and K-NN. Moreover, Table 10 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 7–9 confirm that the proposed fuzzy-rough MRMS method selects a set of features having 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 select a reduced set of features having maximum relevance and significance.

5. Conclusion and future direction

The dimensionality reduction by attribute selection is one of the important problems in pattern recognition, machine learning, and data mining, particularly given the explosive growth of available information. In this regard, the contribution of this paper is three fold, namely,

1. development of a new feature selection method, integrating judiciously the theory of fuzzy-rough sets and merits of the MRMS criterion;
2. application of the proposed method in selecting discriminative and significant features from high dimensional benchmark and microarray gene expression data sets; and
3. compare the performance of the proposed method and some existing methods using the predictive accuracy of three classifiers, namely, nearest neighbor rule, decision tree, and support vector machine.

The proposed method uses the concept of fuzzy-rough feature relevance and significance for finding significant and relevant features of real valued data sets. This formulation is geared towards maximizing the utility of fuzzy-rough sets, feature selection, and the MRMS criterion with respect to knowledge discovery tasks. Through these investigations and experiments, the potential utility of fuzzy-rough sets and the MRMS criterion for attribute selection is demonstrated.

The results obtained on different benchmark and microarray data sets demonstrate that a feature extraction technique such as PCA, ICA, or LDA may provide a reacher feature subset than that obtained using a feature selection algorithm with a higher cost. However, it is very difficult to decide whether to select a feature from original measurement space or extract a new feature by transforming the existing features for a given data set. Hence, a dimensionality reduction algorithm needs to be formulated in near future that can simultaneously select or extract features depending upon the criteria, integrating the merits of both feature selection and extraction techniques.

Acknowledgments

This work is partially supported by the Indian National Science Academy, New Delhi (grant no. SP/YSP/68/2012). The work was done when one of the authors, P. Garai, was a DST-INSPIRE Fellow of the Department of Science and Technology, Government of India.

References

[1] P.A. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach, Prentice Hall, Englewood Cliffs, 1982.
[2] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification and Scene Analysis, John Wiley & Sons, Inc., New York, 1999.
[3] A. Pal, S.K. Pal, Pattern recognition: evolution of methodologies and data mining, in: S.K. Pal, A. Pal (Eds.), Pattern Recognition: from Classical to Modern Approaches, World Scientific, Singapore, 2001, pp. 1–23.
[4] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.
[5] M. Dash, H. Liu, Unsupervised feature selection, in: Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining, 2000, pp. 110–121.
[6] S.K. Pal, P. Mitra, Pattern Recognition Algorithms for Data Mining, CRC Press, Boca Raton, FL, 2004.
[7] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Network 5 (4) (1994) 537–550.
[8] N. Kwak, C.-H. Choi, Input feature selection by mutual information based on Parzen window, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (12) (2002) 1667–1671.
[9] D. Huang, T.W.S. Chow, Effective feature selection scheme using mutual information, Neurocomputing 63 (2004) 325–343.
[10] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, Journal of Bioinformatics and Computational Biology 3 (2) (2005) 185–205.
[11] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1226–1238.
[12] M.A. Hall, Correlation based feature selection for discrete and numerical class machine learning, in: Proceedings of the 17th International Conference on Machine Learning, 2000.
[13] S.K. Das, Feature selection with a linear dependence measure, IEEE Transactions on Computers 20 (9) (1971) 1106–1109.
[14] R.P. Heydorn, Redundancy in feature extraction, IEEE Transactions on Computers (1971) 1051–1054.
[15] M. Bressan, J. Vitria, On the selection and classification of independent features, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (10) (2003) 1312–1317.
[16] P. Maji, S. Paul, Rough sets for selection of molecular descriptors to predict biological activity of molecules, IEEE Transactions on System, Man, and Cybernetics, Part C: Applications and Reviews 40 (6) (2010) 639–648.
[17] P. Maji, S. Paul, Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data, International Journal of Approximate Reasoning 52 (3) (2011) 408–426.
[18] P. Maji, S. Paul, Rough set based feature selection: criteria of Max-Dependency, Max-Relevance, and Max-Significance, in: A. Skowron, Z. Suraj (Eds.), Rough Sets and Intelligent Systems: Professor Zdzislaw Pawlak in Memoriam, Springer, Berlin Heidelberg, 2012, pp. 393–418.
[19] A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, New York, 2001.
[20] P. Maji, S.K. Pal, Feature selection using f-information measures in fuzzy approximation spaces, IEEE Transactions on Knowledge and Data Engineering 22 (6) (2010) 854–867.
[21] M. Dash, H. Liu, Consistency based search in feature selection, Artificial Intelligence 151 (1–2) (2003) 155–176.
[22] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (1979) 224–227.
[23] J.C. Bezdek, N.R. Pal, Some new indexes for cluster validity, IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics 28 (1988) 301–315.
[24] S.K. Pal, R.K. De, J. Basak, Unsupervised feature evaluation: a neuro-fuzzy approach, IEEE Transactions on Neural Network 11 (2) (2000) 366–376.
[25] Z. Pawlak, Rough Sets: Theoretical Aspects of Resoning About Data, Kluwer, Dordrecht, The Netherlands, 1991.
[26] P. Maji, S.K. Pal, RFCM: a hybrid clustering algorithm using rough and fuzzy sets, Fundamenta Informaticae 80 (4) (2007) 475–496.
[27] P. Maji, S.K. Pal, Rough-fuzzy C-medoids algorithm and selection of bio-basis for amino acid sequence analysis, IEEE Transactions on Knowledge and Data Engineering 19 (6) (2007) 859–872.
[28] P. Maji, S.K. Pal, Rough set based generalized fuzzy C-means algorithm and quantitative indices, IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics 37 (6) (2007) 1529–1540.
[29] R. Jensen, Q. Shen, Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches, Wiley-IEEE Press, Hoboken, NJ, 2008.
[30] P. Maji, S.K. Pal, Maximum class separability for rough-fuzzy C-means based brain MR image segmentation, LNCS Transactions on Rough Sets 9 (2008) 114–134.
[31] P. Maji, S.K. Pal, Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging, Wiley-IEEE Computer Society Press, Hoboken, NJ, 2012.
[32] P. Maji, Fuzzy-rough supervised attribute clustering algorithm and classification of microarray data, IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics 41 (1) (2011) 222–233.

[33] P. Maji, S.K. Pal, Fuzzy-rough sets for information measures and selection of relevant genes from microarray data, IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics 40 (3) (2010) 741–752.
[34] S.K. Pal, S. Meher, S. Dutta, Class-dependent rough-fuzzy granular space, dispersion index and classification, Pattern Recognition 45 (7) (2012) 2690–2707.
[35] A. Chouchoulas, Q. Shen, Rough set-aided keyword reduction for text categorisation, Applied Artificial Intelligence 15 (9) (2001) 843–873.
[36] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approach, IEEE Transactions on Knowledge and Data Engineering 16 (12) (2004) 1457–1471.
[37] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, International Journal of General Systems 17 (1990) 191–209.
[38] Q. Hu, D. Yu, Z. Xie, J. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Transactions on Fuzzy Systems 14 (2) (2007) 191–201.
[39] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Information Sciences 178 (2008) 3577–3594.
[40] S.K. Pal, P.K. Pramanik, Fuzzy measures in determining seed points in clustering, Pattern Recognition Letters 4 (3) (1986) 159–164.
[41] S.K. Pal, S. Mitra, Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing, John Wiley & Sons, New York, 1999.
[42] J.R. Quinlan, Induction of decision trees, Machine Learning 1 (1986) 81–106.
[43] K. Kira, L.A. Rendell, The feature selection problem: traditional methods and a new algorithm, in: Proceedings of the 10th National Conference on Artificial Intelligence, MIT Press, 1992, pp. 129–134.
[44] R. Gilad-Bachrach, A. Navot, N. Tishby, Margin based feature selection: theory and algorithms, in: Proceedings of the 21st International Conference on Machine Learning, 2004.
[45] P. Maji, *f*-Information measures for efficient selection of discriminative genes from microarray data, IEEE Transactions on Biomedical Engineering 56 (4) (2009) 1063–1069.
[46] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[47] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, CA, 1993.
[48] A. Frank, A. Asuncion, UCI Machine Learning Repository, 2010 http://archive.ics.uci.edu/ml
[49] Kent Ridge Bio-medical Data Set Repository. http://sdmc.lit.org.sg/GEDatasets/Datasets.html
[50] A. Aspin, Tables for use in comparisons whose accuracy involves two variances, Biometrika 36 (1949) 245–271.