Feature Selection Using f-Information Measures in Fuzzy Approximation Spaces

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

Abstract—The selection of nonredundant and relevant features of real-valued data sets is a highly challenging problem. A novel feature selection method is presented here based on fuzzy-rough sets by maximizing the relevance and minimizing the redundancy of the selected features. By introducing the fuzzy equivalence partition matrix, a novel representation of Shannon's entropy for fuzzy approximation spaces is proposed to measure the relevance and redundancy of features suitable for real-valued data sets. The fuzzy equivalence partition matrix also offers an efficient way to calculate many more information measures, termed as f-information measures. Several f-information measures are shown to be effective for selecting nonredundant and relevant features of real-valued data sets. This paper compares the performance of different f-information measures for feature selection in fuzzy approximation spaces. Some quantitative indexes are introduced based on fuzzy-rough sets for evaluating the performance of proposed method. The effectiveness of the proposed method, along with a comparison with other methods, is demonstrated on a set of real-life data sets.

Index Terms—Pattern recognition, data mining, feature selection, fuzzy-rough sets, f-information measures.

1 Introduction

Feature selection or dimensionality reduction of a data set is an essential preprocessing step used for pattern recognition, data mining, machine learning, etc., [1], [2]. It is an important problem related to mining large data sets, both in dimension and size. Prior to analysis of the data set, preprocessing the data to obtain a smaller set of representative features and retaining the optimal salient characteristics of the data not only decrease the processing time, but also lead to more compactness of the models learned and better generalization. Hence, the general criterion for reducing the dimension is to preserve most relevant information of the original data according to some optimality criteria [1], [2].

Conventional methods of feature selection involve evaluating different feature subsets using some index and selecting the best among them. An optimal feature subset is always relative to a certain criterion. In general, different criteria may lead to different optimal feature subset. However, every criterion tries to measure the discriminating ability of a feature or a subset of features to distinguish the different class labels. While the distance measure is a very traditional discrimination or divergence measure, the dependence or correlation measure is mainly utilized to find the correlation between two features or a feature and a class [3]. As these two measures depend on the actual values of the training data, they are very much sensitive to the noise or outlier of the data set. On the other hand, the information measures, such as the entropy and mutual information [4], compute the amount of information or the uncertainty of a feature for classification. As the information measure depends only on the probability distribution of a random variable rather than on its actual values, it has been widely used in feature selection [4], [5].

Information measures are defined as the measures of the distance between a joint probability distribution and the product of the marginal distributions [6]. They constitute a subclass of the divergence measures, which are measures of the distance between two arbitrary distributions. A specific class of information (divergence) measures, of which mutual information is a member, is formed by the f-information (f-divergence) measures [6], [7]. Several f-information measures have been successfully used in medical image registration [7] and gene selection [8] problems, and shown to yield significantly more accurate results than mutual information.

Rough set theory [9] is a new paradigm to deal with uncertainty, vagueness, and incompleteness. It has been applied to fuzzy rule extraction, reasoning with uncertainty, fuzzy modeling, classification, feature selection, etc., [9], [10]. However, there are usually real-valued data and fuzzy information in real-world applications. Combining fuzzy and rough sets provides an important direction in reasoning with uncertainty for real-valued data sets [10], [11], [12]. Both fuzzy and rough sets provide a mathematical framework to capture uncertainties associated with the data [12]. They are complementary in some aspects. The generalized theories of rough-fuzzy and fuzzy-rough sets have been applied successfully to feature selection of real-valued data [10], fuzzy decision rule extraction, rough-fuzzy clustering [11], [13], etc.

In [10], Jensen and Shen introduced the fuzzy-rough quick reduct algorithm for feature selection of real-valued data sets. In [14], Hu et al. have used the concept of fuzzy equivalence relation matrix to compute entropy and mutual information in fuzzy approximation spaces, which can be used for feature selection of real-valued data sets. However, many useful information measures such as several f-information measures cannot be computed from the fuzzy equivalence relation matrix introduced in [14] as

• *The authors are with the Machine Intelligence Unit, Indian Statistical Institute, 203 B T Road, Kolkata, India.*
E-mail: {pmaji, sankar}@isical.ac.in.

Manuscript received 14 Jan. 2008; revised 22 Jan. 2009; accepted 7 Apr. 2009; published online 1 May 2009.
Recommended for acceptance by B.C. Ooi
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2008-01-0023.

it does not provide a way to compute marginal and joint distributions directly. Also, the fuzzy-rough-set-based feature selection methods proposed in [10], [14] select the relevant features of a data set without considering the redundancy among them.

In this paper, a novel feature selection method is proposed, which employs fuzzy-rough sets to provide a means by which discrete- or real-valued noisy data (or a mixture of both) can be effectively reduced without the need for user-specified information. Moreover, the proposed method can be applied to data with continuous or nominal decision attributes, and can be applied to regression as well as classification data sets. The proposed method selects a subset of features from the whole feature set by maximizing the relevance and minimizing the redundancy of the selected features. The relevance and redundancy of the features are calculated using the f-information measures in fuzzy approximation spaces. Using the concept of fuzzy equivalence partition matrix, the f-information measures are calculated for both condition and decision attributes. Hence, the only information required in the proposed feature selection method is in the form of fuzzy partitions for each attribute, which can be automatically derived from the given data set. Several quantitative measures are introduced based on fuzzy-rough sets to evaluate the performance of the proposed feature selection method. The effectiveness of the proposed method, along with a comparison with other methods, is demonstrated on a set of real-life data.

The structure of the rest of this paper is as follows: Section 2 briefly introduces the necessary notions of rough sets and fuzzy-rough sets. In Section 3, the formulas of Shannon's entropy are introduced for fuzzy approximation spaces with a fuzzy equivalence partition matrix. The f-information measures for fuzzy approximation spaces are presented next in Section 4. The proposed feature selection method based on f-information measures for fuzzy approximation spaces is described in Section 5. Several quantitative measures are presented in Section 6 to evaluate the performance of the proposed method. A few case studies and a comparison with other methods are presented in Section 7. Concluding remarks are given in Section 8.

2 Rough Sets and Fuzzy-Rough Sets

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

2.1 Rough Sets

The theory of rough sets begins with the notion of an approximation space, which is a pair $<\mathbb{U}, \mathbb{A}>$, where $\mathbb{U}$ be a nonempty set (the universe of discourse), $\mathbb{U} = \{x_1, \ldots, x_i, \ldots, x_n\}$, and $\mathbb{A}$ is a family of attributes, also called knowledge in the universe. V is the value domain of $\mathbb{A}$ and $\hat{f}$ is an information function $\hat{f} : \mathbb{U} \times \mathbb{A} \rightarrow V$. An approximation space is also called an information system [9].

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

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

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

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

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

Given an arbitrary set $X \subseteq \mathbb{U}$, in general, it may not be possible to describe X precisely in $<\mathbb{U}, \mathbb{A}>$. One may characterize X by a pair of lower and upper approximations defined as follows [9]:

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

That is, 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 $<\underline{\mathbb{P}}(X), \overline{\mathbb{P}}(X)>$ is the representation of an ordinary set X in the approximation space $<\mathbb{U}, \mathbb{A}>$ 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 $<\mathbb{U}, \mathbb{A}>$ iff $\underline{\mathbb{P}}(X) = \overline{\mathbb{P}}(X)$. Otherwise, X is indefinable and termed as a rough set. $BN_{\mathbb{P}}(X) = \overline{\mathbb{P}}(X) \setminus \underline{\mathbb{P}}(X)$ is called a boundary set.

An information system $<\mathbb{U}, \mathbb{A}>$ 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

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|POS_{\mathbb{C}}(\mathbb{D})|}{|\mathbb{U}|}, \quad (3)$$

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

2.2 Fuzzy-Rough Sets

A crisp equivalence relation induces a crisp partition of the universe and generates a family of crisp equivalence classes. Correspondingly, a fuzzy equivalence relation generates a fuzzy partition of the universe and a series of fuzzy equivalence classes, which are also called fuzzy knowledge granules. This means that the decision and condition attributes may all be fuzzy [10], [12].

Let $<\mathbb{U}, \mathbb{A}>$ 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 [12]:

$$\mu_{\underline{\mathbb{P}}X}(F_i) = \inf_x \{\max\{(1 - \mu_{F_i}(x)), \mu_X(x)\}\} \quad \forall i, \quad (4)$$

$$\mu_{\overline{\mathbb{P}}X}(F_i) = \sup_x \{\min\{\mu_{F_i}(x), \mu_X(x)\}\} \quad \forall i, \quad (5)$$

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. Note that although the universe of discourse in feature selection is finite, this is not the case, in general, hence the use of sup and inf. These

definitions diverge a little from the crisp upper and lower approximations, as the memberships of individual objects to the approximations are not explicitly available. As a result of this, the fuzzy lower and upper approximations can be defined as [10]

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

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

The tuple $<\underline{\mathbb{P}}X, \overline{\mathbb{P}}X>$ 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

$$\mu_{POS_{\mathbb{C}}(\mathbb{D})}(x) = \sup_{X \in \mathbb{U}/\mathbb{D}} \mu_{\underline{\mathbb{C}}X}(x), \tag{8}$$

where $\mathbb{A} = \mathbb{C} \cup \mathbb{D}$. Using the definition of fuzzy positive region, the dependency function can be defined as follows [10]:

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{|\mu_{POS_{\mathbb{C}}(\mathbb{D})}(x)|}{|\mathbb{U}|} = \frac{1}{|\mathbb{U}|} \sum_{x \in \mathbb{U}} \mu_{POS_{\mathbb{C}}(\mathbb{D})}(x). \tag{9}$$

3 Information Measure on Fuzzy Approximation Spaces

In this section, the Shannon's information measure [15] is introduced to compute the knowledge quantity of a fuzzy attribute set or a fuzzy partition of $\mathbb{U}$. Shannon's information entropy [15] just works in the case where a crisp equivalence relation or a crisp partition is defined. That is, it is suitable for Pawlak's approximation space [9]. In this section, a novel formula to compute Shannon's entropy with a fuzzy equivalence partition matrix is presented, which will be used to measure the information on fuzzy approximation spaces.

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

$$\mathbb{M}_{\mathbb{A}} = \begin{pmatrix} m_{11}^{\mathbb{A}} & m_{12}^{\mathbb{A}} & \cdots & m_{1n}^{\mathbb{A}} \\ m_{21}^{\mathbb{A}} & m_{22}^{\mathbb{A}} & \cdots & m_{2n}^{\mathbb{A}} \\ \cdots & \cdots & \cdots & \cdots \\ m_{c1}^{\mathbb{A}} & m_{c2}^{\mathbb{A}} & \cdots & m_{cn}^{\mathbb{A}} \end{pmatrix}, \tag{10}$$

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

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

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

Using the concept of fuzzy equivalence partition matrix, the dependency between condition attribute set $\mathbb{C}$ and decision attribute set $\mathbb{D}$ can be redefined as follows:

$$\gamma_{\mathbb{C}}(\mathbb{D}) = \frac{1}{n} \sum_{j=1}^{n} \kappa_j, \tag{11}$$

where $\mathbb{C} \cup \mathbb{D} = \mathbb{A}$ and

$$\kappa_j = \sup_k \left\{\sup_i \left\{\min\left\{m_{ij}^{\mathbb{C}}, \inf_l \left\{\max\{1 - m_{il}^{\mathbb{C}}, m_{kl}^{\mathbb{D}}\}\right\}\right\}\right\}\right\}. \tag{12}$$

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

$$F_i = \left\{m_{i1}^{\mathbb{A}}/x_1 + m_{i2}^{\mathbb{A}}/x_2 + \cdots + m_{in}^{\mathbb{A}}/x_n\right\}. \tag{13}$$

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

$$|F_i| = \sum_{j=1}^{n} m_{ij}^{\mathbb{A}}, \tag{14}$$

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

$$H(\mathbb{A}) = -\sum_{i=1}^{c} \lambda_{F_i} \log \lambda_{F_i}, \tag{15}$$

where $\lambda_{F_i} = \frac{|F_i|}{n}$, called a fuzzy relative frequency, and c is the number of fuzzy equivalence partitions or classes. The measure $H(\mathbb{A})$ has the same form as the Shannon's entropy [15]. The information quantity or the entropy value increases monotonously with the discernibility power of the fuzzy attributes.

Given $<\mathbb{U}, \mathbb{A}>$, $\mathbb{P}$ and $\mathbb{Q}$ are two subsets of fuzzy attribute set $\mathbb{A}$. The information quantity corresponding to $\mathbb{P}$ and $\mathbb{Q}$ is given by

$$H(\mathbb{P}) = -\sum_{i=1}^{p} \lambda_{P_i} \log \lambda_{P_i}, \tag{16}$$

$$H(\mathbb{Q}) = -\sum_{j=1}^{q} \lambda_{Q_j} \log \lambda_{Q_j}, \tag{17}$$

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

$$H(\mathbb{P}\mathbb{Q}) = -\sum_{k=1}^{r} \lambda_{R_k} \log \lambda_{R_k}, \tag{18}$$

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

$$\lambda_{R_k} = \lambda_{P_iQ_j} = \frac{|P_i \cap Q_j|}{n}; \quad \text{where } k = (i-1)q + j. \tag{19}$$

That is, the joint frequency λ_{R_k} can be calculated from the $r \times n$ fuzzy equivalence partition matrix $\mathbb{M}_{\mathbb{PQ}}$, where

$$\mathbb{M}_{\mathbb{PQ}} = \mathbb{M}_{\mathbb{P}} \cap \mathbb{M}_{\mathbb{Q}} \quad \text{and} \quad m_{kl}^{\mathbb{PQ}} = m_{il}^{\mathbb{P}} \cap m_{jl}^{\mathbb{Q}}. \tag{20}$$

Similarly, the conditional entropy of $\mathbb{P}$ conditioned to $\mathbb{Q}$ is defined as

$$H(\mathbb{P}|\mathbb{Q}) = -\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{|P_i \cap Q_j|}{n} \log \frac{|P_i \cap Q_j|}{|Q_j|} \tag{21}$$

$$= -\sum_{i=1}^{p}\sum_{j=1}^{q} \left\{ \frac{|P_i \cap Q_j|}{n} \log \frac{|P_i \cap Q_j|}{n} - \frac{|P_i \cap Q_j|}{n} \log \frac{|Q_j|}{n} \right\}$$

$$= -\left\{ \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{|P_i \cap Q_j|}{n} \log \frac{|P_i \cap Q_j|}{n} - \sum_{j=1}^{q} \frac{|Q_j|}{n} \log \frac{|Q_j|}{n} \right\}.$$

That is, the conditional entropy of $\mathbb{P}$ conditioned to $\mathbb{Q}$ is

$$H(\mathbb{P}|\mathbb{Q}) = -\sum_{k=1}^{r} \lambda_{R_k} \log \lambda_{R_k} + \sum_{j=1}^{q} \lambda_{Q_j} \log \lambda_{Q_j}, \tag{22}$$

where

$$\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{|P_i \cap Q_j|}{n} = \sum_{j=1}^{q} \frac{|Q_j|}{n} \quad \text{and} \quad \lambda_{Q_j} = \frac{|Q_j|}{n}.$$

Thus,

$$H(\mathbb{P}|\mathbb{Q}) = H(\mathbb{PQ}) - H(\mathbb{Q}). \tag{23}$$

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

$$I(\mathbb{PQ}) = H(\mathbb{P}) - H(\mathbb{P}|\mathbb{Q}) = H(\mathbb{P}) + H(\mathbb{Q}) - H(\mathbb{PQ}). \tag{24}$$

The mutual information $I(\mathbb{PQ})$ between two fuzzy attribute sets $\mathbb{P}$ and $\mathbb{Q}$ quantifies the information shared by both of them. If $\mathbb{P}$ and $\mathbb{Q}$ do not share much information, the value of $I(\mathbb{PQ})$ between them is small. While two highly nonlinearly correlated attribute sets will demonstrate a high mutual information value. The attribute sets can be both the condition attributes and the decision attributes in this study. The necessity for a fuzzy condition attribute to be an independent and informative feature can, therefore, be determined by the shared information between this attribute and the rest as well as the shared information between this attribute and the decision attribute.

4 f-Information Measures and Fuzzy Approximation Spaces

The extent to which two probability distributions differ can be expressed by a so-called measure of divergence. Such a measure will reach a minimum value when two probability distributions are identical and the value increases with increasing disparity between two distributions. A specific class of divergence measures is the set of f-divergence [6]. For two discrete probability distributions $P = \{p_i | i = 1, \ldots, n\}$ and $Q = \{q_i | i = 1, \ldots, n\}$, the f-divergence is defined as

$$f(P\|Q) = \sum_i q_i f\left(\frac{p_i}{q_i}\right). \tag{25}$$

A special case of f-divergence measures is the f-information measures. These are defined similarly to f-divergence measures, but apply only to specific probability distributions, namely, the joint probability of two variables and their marginal probabilities' product. Thus, f-information is a measure of dependence: it measures the distance between a given joint probability and joint probability when variables are independent [6], [7].

In this section, several frequently used f-information is reported for fuzzy approximation spaces based on the concept of fuzzy relative frequency. The f-information measures in fuzzy approximation spaces calculate the distance between a given joint frequency $\lambda_{R_k} (= \lambda_{P_iQ_j})$ and the joint frequency when the variables are independent $(\lambda_{P_i}\lambda_{Q_j})$. In the following analysis, it is assumed that all frequency distributions are complete, that is, $\sum \lambda_{P_i} = \sum \lambda_{Q_j} = \sum \lambda_{P_iQ_j} = 1$.

4.1 V-Information

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

$$V(R\|P \times Q) = \sum_{i,j,k} |\lambda_{R_k} - \lambda_{P_i}\lambda_{Q_j}|, \tag{26}$$

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

4.2 I_α-Information

The I_α-information can be defined as follows:

$$I_\alpha(R\|P \times Q) = \frac{1}{\alpha(\alpha-1)} \left(\sum_{i,j,k} \frac{(\lambda_{R_k})^\alpha}{(\lambda_{P_i}\lambda_{Q_j})^{\alpha-1}} - 1 \right), \tag{27}$$

for $\alpha \neq 0, \alpha \neq 1$. The class of I_α-information includes mutual information, which equals I_α for the limit $\alpha \to 1$, that is,

$$I_1(R\|P \times Q) = \sum_{i,j,k} \lambda_{R_k} \log\left(\frac{\lambda_{R_k}}{\lambda_{P_i}\lambda_{Q_j}} \right). \tag{28}$$

4.3 M_α-Information

The M_α-information is defined [6], [7] as follows:

$$M_\alpha(x) = |x^\alpha - 1|^{\frac{1}{\alpha}}, \quad 0 < \alpha \leq 1. \tag{29}$$

When applying this function in the definition of an f-information measure on fuzzy approximation spaces, the resulting M_α-information measures are

$$M_\alpha(R\|P \times Q) = \sum_{i,j,k} |(\lambda_{R_k})^\alpha - (\lambda_{P_i}\lambda_{Q_j})^\alpha|^{\frac{1}{\alpha}}, \tag{30}$$

for $0 < \alpha \leq 1$. These constitute a generalized version of V-information. That is, the M_α-information is identical to V-information for $\alpha = 1$.

4.4 χ^α-Information

The class of χ^α-information measures, proposed by Liese [6], [7], is as follows:

$$\chi^\alpha(x) = \begin{cases} |1 - x^\alpha|^{\frac{1}{\alpha}}, & \text{for } 0 < \alpha \leq 1, \\ |1 - x|^\alpha, & \text{for } \alpha > 1. \end{cases} \tag{31}$$

For $0 < \alpha \leq 1$, this function equals to the M_α function. The χ^α and M_α-information measures are, therefore, also identical for $0 < \alpha \leq 1$. For $\alpha > 1$, χ^α-information can be written as

$$\chi^\alpha(R \| P \times Q) = \sum_{i,j,k} \frac{|\lambda_{R_k} - \lambda_{P_i}\lambda_{Q_j}|^\alpha}{(\lambda_{P_i}\lambda_{Q_j})^{\alpha-1}}. \tag{32}$$

4.5 Renyi Distance

The Renyi distance, a measure of information of order α [6], [7], can be defined as

$$\mathcal{R}_\alpha(R \| P \times Q) = \frac{1}{\alpha - 1} \log \sum_{i,j,k} \frac{(\lambda_{R_k})^\alpha}{(\lambda_{P_i}\lambda_{Q_j})^{\alpha-1}}, \tag{33}$$

for $\alpha \neq 0, \alpha \neq 1$. It reaches its minimum value when λ_{R_k} and $(\lambda_{P_i}\lambda_{Q_j})$ are identical, in which case the summation reduces to $\sum \lambda_{R_k}$. As we assume complete frequency distributions, the sum is 1 and the minimum value of the measure is, therefore, equal to zero. The limit of Renyi's measure for α approaching 1 equals $I_1(R \| P \times Q)$, which is the mutual information.

5 Proposed Feature Selection Method

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

5.1 Feature Selection Using f-Information

Let $\mathbb{C} = \{\mathbb{C}_1, \ldots, \mathbb{C}_i, \ldots, \mathbb{C}_j, \ldots, \mathbb{C}_\mathcal{D}\}$ denotes the set of condition attributes or features of a given data set and $\mathbb{S}$ be the set of selected features. Define $\tilde{f}(\mathbb{C}_i, \mathbb{D})$ as the relevance of the fuzzy condition attribute $\mathbb{C}_i$ with respect to the fuzzy decision attribute $\mathbb{D}$, while $\tilde{f}(\mathbb{C}_i, \mathbb{C}_j)$ as the redundancy between two fuzzy condition attributes $\mathbb{C}_i$ and $\mathbb{C}_j$. The total relevance of all selected features is, therefore, given by

$$\mathcal{J}_{\text{relev}} = \sum_{\mathbb{C}_i \in \mathbb{S}} \tilde{f}(\mathbb{C}_i, \mathbb{D}), \tag{34}$$

while total redundancy among the selected features is

$$\mathcal{J}_{\text{redun}} = \sum_{\mathbb{C}_i, \mathbb{C}_j \in \mathbb{S}} \tilde{f}(\mathbb{C}_i, \mathbb{C}_j). \tag{35}$$

Therefore, the problem of selecting a set $\mathbb{S}$ of nonredundant and relevant features from the whole set of condition features $\mathbb{C}$ is equivalent to maximize $\mathcal{J}_{\text{relev}}$ and minimize $\mathcal{J}_{\text{redun}}$, that is, to maximize the objective function $\mathcal{J}$, where

$$\mathcal{J} = \mathcal{J}_{\text{relev}} - \beta \mathcal{J}_{\text{redun}} = \sum_i \tilde{f}(\mathbb{C}_i, \mathbb{D}) - \beta \sum_{i,j} \tilde{f}(\mathbb{C}_i, \mathbb{C}_j), \tag{36}$$

where β is a weight parameter. To solve the above problem, the greedy algorithm of Battiti [4] is used that follows next.

1. Initialize $\mathbb{C} \leftarrow \{\mathbb{C}_1, \ldots, \mathbb{C}_i, \ldots, \mathbb{C}_j, \ldots, \mathbb{C}_\mathcal{D}\}, \mathbb{S} \leftarrow \emptyset$.
2. Generate fuzzy equivalence partition matrix for each condition and decision attribute.
3. Calculate the relevance value $\tilde{f}(\mathbb{C}_i, \mathbb{D})$ of each feature $\mathbb{C}_i \in \mathbb{C}$.
4. Select feature $\mathbb{C}_i$ as the first feature that has the highest relevance $\tilde{f}(\mathbb{C}_i, \mathbb{D})$. In effect, $\mathbb{C}_i \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathbb{C}_i$.
5. Generate resultant equivalence partition matrix between selected features and each of remaining features of $\mathbb{C}$.
6. Calculate the redundancy between selected features of $\mathbb{S}$ and each of remaining features of $\mathbb{C}$.
7. From the remaining features of $\mathbb{C}$, select feature $\mathbb{C}_j$ that maximizes

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

   As a result of that, $\mathbb{C}_j \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus \mathbb{C}_j$.
8. Repeat the above three steps until the desired number of features is selected.

The relevance of a fuzzy condition attribute with respect to the fuzzy decision attribute and the redundancy between two fuzzy condition attributes can be calculated using any one of f-information measures on fuzzy approximation spaces.

5.2 Computational Complexity

The f-information-measure-based proposed feature selection method has low computational complexity with respect to both number of features and number of samples or objects of the original data set. Prior to computing the relevance or redundancy of a fuzzy condition attribute, the fuzzy equivalence partition matrix for each condition and decision attribute is to be generated first. The computational complexity to generate a $(c \times n)$ fuzzy equivalence partition matrix is $\mathcal{O}(cn)$, where c represents the number of fuzzy equivalence partitions and n is the total number of objects in the data set. However, two fuzzy equivalence partition matrices with size $(p \times n)$ and $(r \times n)$ have to be generated to compute the relevance of a fuzzy condition attribute with respect to the fuzzy decision attribute, where p and r represent the number of fuzzy equivalence partitions of fuzzy condition attribute

and fuzzy decision attribute, respectively. Hence, the total time complexity to calculate the relevance of a fuzzy condition attribute using any one of the f-information measures is $(\mathcal{O}(pn) + \mathcal{O}(rn) + \mathcal{O}(prn)) = \mathcal{O}(prn)$. Similarly, the complexity to calculate the redundancy between two fuzzy condition attributes with p and q number of fuzzy equivalence partitions using any one of the f-information measures is $\mathcal{O}(pqn)$. Hence, the overall time complexity to calculate both relevance and redundancy of a fuzzy condition attribute is $(\mathcal{O}(prn) + \mathcal{O}(pqn)) = \mathcal{O}(n)$ as $p, q, r << n$. In effect, the selection of a set of d nonredundant and relevant features from the whole set of $\mathcal{D}$ features using the proposed first order incremental search method has an overall computational complexity of $\mathcal{O}(nd\mathcal{D})$.

5.3 Fuzzy Equivalence Classes

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]. In the proposed feature selection method, 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 $\pi(x; \bar{c}, \sigma)$ [16] represents a set of points clustered around $\bar{c}$, where

$$\pi(x;\bar{c},\sigma) = \begin{cases} 2\left(1 - \frac{\|x-\bar{c}\|}{\sigma}\right)^2, & \text{for } \frac{\sigma}{2} \le \|x-\bar{c}\| \le \sigma, \\ 1 - 2\left(\frac{\|x-\bar{c}\|}{\sigma}\right)^2, & \text{for } 0 \le \|x-\bar{c}\| \le \frac{\sigma}{2}, \\ 0, & \text{otherwise}, \end{cases} \tag{37}$$

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 $\|x - \bar{c}\| = (\frac{\sigma}{2})$, the membership value of x is 0.5 and this is called a crossover point [16].

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.

5.3.1 Choice of Parameters of π Function

The parameters $\bar{c}$ and σ of each π fuzzy set are computed according to the procedure reported in [16]. Let $\bar{m}_i$ be the mean of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$ along the ith feature $\mathbb{C}_i$. 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 range $[\mathbb{C}_{i_{\min}}, \bar{m}_i)$ and $(\bar{m}_i, \mathbb{C}_{i_{\max}}]$, respectively, where $\mathbb{C}_{i_{\max}}$ and $\mathbb{C}_{i_{\min}}$ denote the upper and lower bounds of the dynamic range of feature $\mathbb{C}_i$ for the training set. For three fuzzy sets low, medium, and high, the centers and corresponding radii are as follows [16]:

$$\bar{c}_{\text{low}}(\mathbb{C}_i) = \bar{m}_{i_l}; \bar{c}_{\text{medium}}(\mathbb{C}_i) = \bar{m}_i; \bar{c}_{\text{high}}(\mathbb{C}_i) = \bar{m}_{i_h}, \tag{38}$$

$$\begin{aligned} \sigma_{\text{low}}(\mathbb{C}_i) &= 2(\bar{c}_{\text{medium}}(\mathbb{C}_i) - \bar{c}_{\text{low}}(\mathbb{C}_i)), \\ \sigma_{\text{high}}(\mathbb{C}_i) &= 2(\bar{c}_{\text{high}}(\mathbb{C}_i) - \bar{c}_{\text{medium}}(\mathbb{C}_i)), \\ \sigma_{\text{medium}}(\mathbb{C}_i) &= \eta \times \frac{\mathrm{A}}{\mathrm{B}}, \end{aligned} \tag{39}$$

where

$$\mathrm{A} = \{\sigma_{\text{low}}(\mathbb{C}_i)(\mathbb{C}_{i_{\max}} - c_{\text{medium}}(\mathbb{C}_i)) + \sigma_{\text{high}}(\mathbb{C}_i)(c_{\text{medium}}(\mathbb{C}_i) - \mathbb{C}_{i_{\min}})\}; \mathrm{B} = \{\mathbb{C}_{i_{\max}} - \mathbb{C}_{i_{\min}}\},$$

where η is a multiplicative parameter controlling the extent of the overlapping. The distribution of the patterns or objects along each feature axis is taken into account, while computing the corresponding centers and radii of the 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 or patterns.

5.3.2 Fuzzy Equivalence Partition Matrix

The $c \times n$ fuzzy equivalence partition matrix $\mathbb{M}_{\mathbb{C}_i}$, corresponding to the ith feature $\mathbb{C}_i$, can be calculated from the c-fuzzy equivalence classes of the objects $x = \{x_1, \ldots, x_j, \ldots, x_n\}$, where

$$m_{kj}^{\mathbb{C}_i} = \frac{\pi(x_j; \bar{c}_k, \sigma_k)}{\sum_{l=1}^{c} \pi(x_j; \bar{c}_l, \sigma_l)}. \tag{40}$$

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

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

In effect, each position $m_{kj}^{\mathbb{C}_i}$ of the fuzzy equivalence partition matrix $\mathbb{M}_{\mathbb{C}_i}$ must satisfy the following conditions:

$$m_{kj}^{\mathbb{C}_i} \in [0,1]; \sum_{k=1}^{c} m_{kj}^{\mathbb{C}_i} = 1, \forall j \text{ and for any value of } k, \text{if}$$
$$s = \arg\max_j \left\{m_{kj}^{\mathbb{C}_i}\right\}, \text{then } \max_j \left\{m_{kj}^{\mathbb{C}_i}\right\} = \max_l \left\{m_{ls}^{\mathbb{C}_i}\right\} > 0.$$

6 Quantitative Measures

In this section, two new quantitative indexes are presented, along with some existing indexes, to evaluate the performance of proposed method. The proposed two indexes are based on the concept of fuzzy-rough sets.

6.1 Fuzzy-Rough-Set-Based Quantitative Indexes

Using the definition of fuzzy positive region, two new indexes are introduced next.

6.1.1 $\mathbb{RELEV}$ Index

The $\mathbb{RELEV}$ index is defined as

$$\mathbb{RELEV} = \frac{1}{|\mathbb{S}|} \sum_{\mathbb{C}_i \in \mathbb{S}} \gamma_{\mathbb{C}_i}(\mathbb{D}), \tag{41}$$

where $\gamma_{\mathbb{C}_i}(\mathbb{D})$ represents the degree of dependency of decision attribute $\mathbb{D}$ on the condition attribute $\mathbb{C}_i$, which can be calculated using (11). That is, $\mathbb{RELEV}$ index is the average relevance of all selected features. A good feature selection algorithm should make all selected features as relevant as possible. The $\mathbb{RELEV}$ index increases with the

increase in relevance of each selected feature. Therefore, for a given data set and number of selected features, the higher the relevance of each selected feature, the higher would be the $\mathbb{RELEV}$ index.

6.1.2 $\mathbb{REDUN}$ Index

It can be defined as

$$\mathbb{REDUN} = \frac{1}{2|\mathbb{S}||\mathbb{S}-1|}\sum_{\mathbb{C}_i,\mathbb{C}_j}\{\gamma_{\mathbb{C}_i}(\mathbb{C}_j)+\gamma_{\mathbb{C}_j}(\mathbb{C}_i)\}, \quad (42)$$

where $\gamma_{\mathbb{C}_i}(\mathbb{C}_j)$ represents the degree of dependency of the condition attribute $\mathbb{C}_j$ on another condition attribute $\mathbb{C}_i$. The $\mathbb{REDUN}$ index calculates the amount of redundancy among the selected features. A good feature selection algorithm should make the redundancy among all selected features as low as possible. The $\mathbb{REDUN}$ index minimizes the redundancy between selected features.

6.2 Existing Feature Evaluation Indexes

Some existing indexes are described next that are used for evaluating the effectiveness of the selected features.

6.2.1 Class Separability

Class separability S of a data set is defined as [2]

$$S = trace\big(S_b^{-1}S_w\big), \quad (43)$$

where S_w and S_b represent the within class and between class scatter matrix, respectively, and defined as follows:

$$S_w = \sum_{j=1}^{C} p_j E\{(X-\mu_j)(X-\mu_j)^T|w_j\} = \sum_{j=1}^{C} p_j\Sigma_j, \quad (44)$$

$$S_b = \sum_{j=1}^{C}(\mu_j - M_0)(\mu_j - M_0)^T; \quad \text{where } M_0 = \sum_{j=1}^{C} p_j\mu_j, \quad (45)$$

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

6.2.2 C4.5 Classification Error

The C4.5 [5] is a popular decision-tree-based classification algorithm. It is used for evaluating the effectiveness of 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 it builds are represented in the form of decision trees, which can be further examined.

6.2.3 K-NN Classification Error

The K-nearest neighbor (K-NN) rule [1] is used for evaluating the effectiveness of the reduced feature set for classification. It classifies samples based on the closest training samples in the feature space. A sample is classified by a majority vote of its K-neighbors, with the sample being assigned to the class most common among its K-nearest neighbors. The value of K, chosen for the K-NN, is the square root of number of samples in training set.

6.2.4 Entropy

Let the distance between two data points x_i and x_j be

$$\mathrm{D}_{ij} = \left[\sum_{k=1}^{d}\left(\frac{x_{ik}-x_{jk}}{max_k - min_k}\right)^2\right]^{\frac{1}{2}}, \quad (46)$$

where x_{ik} denotes feature value for x_i along kth direction, and max_k and min_k are the maximum and minimum values computed over all the samples along kth axis, and d is the number of selected features. Similarity, between x_i and x_j are given by $\mathrm{sim}(i,j) = e^{-\alpha \mathrm{D}_{ij}}$, where α is a positive constant. A possible value of α is $\frac{-\ln 0.5}{\tilde{\mathrm{D}}}$. $\tilde{\mathrm{D}}$ is the average distance between data points computed over the entire data set. Entropy is then defined as [17]:

$$\begin{aligned} E = -\sum_{i=1}^{n}\sum_{j=1}^{n}&(\mathrm{sim}(i,j)\times\log(\mathrm{sim}(i,j)) \\ &+(1-\mathrm{sim}(i,j))\times\log(1-\mathrm{sim}(i,j)). \end{aligned} \quad (47)$$

If the data are uniformly distributed in the feature space, entropy is maximum. When the data have well-formed clusters, uncertainty is low and so is entropy.

6.2.5 Representation Entropy

Let the eigenvalues of the $d\times d$ covariance matrix of a feature set of size d be $\lambda_j, j=1,\ldots,d$. Let

$$\tilde{\lambda}_j = \frac{\lambda_j}{\sum_{j=1}^{d}\lambda_j}, \quad (48)$$

where $\tilde{\lambda}_j$ has the similar properties like probability, namely, $0\le\tilde{\lambda}_j\le 1$ and $\sum_{j=1}^{d}\tilde{\lambda}_j = 1$. Hence, an entropy function can be defined as [2]

$$H_R = -\sum_{j=1}^{d}\tilde{\lambda}_j\log\tilde{\lambda}_j. \quad (49)$$

The function H_R attains a minimum value (zero) when all the eigenvalues except one are zero or, in other words, when all the information is present along a single coordinate direction. If all the eigenvalues are equal, that is, information is equally distributed among all the features, H_R is maximum and so is the uncertainty involved in feature reduction. The above measure is known as representation entropy. Since the proposed method takes into account the redundancy among the selected features, it is expected that the reduced feature set attains a high value of representation entropy.

7 Experimental Results

The performance of the proposed method based on f-information measures is extensively studied. Based on the argumentation given in Section 4, following information measures are chosen to include in the study.

MI:	mutual information;	VI:	V-information;
I_α:	for $\alpha\neq 0$, $\alpha\neq 1$;	M_α:	for $0<\alpha<1$;
χ^α:	for $\alpha>1$;	$\mathcal{R}_\alpha$:	for $\alpha\neq 0$, $\alpha\neq 1$;
C:	fuzzy;	D:	crisp.

TABLE 1
Classification Error of C4.5 for Mutual-Information-Based Feature Selection on Different Data Sets

Data Set	Selected Features	Measure	$\eta = 0.7$ / β			$\eta = 1.1$ / β			$\eta = 1.5$ / β			$\eta = 1.9$ / β		
			0.0	0.5	1.0	0.0	0.5	1.0	0.0	0.5	1.0	0.0	0.5	1.0
E-Coli ($\mathcal{D} = 7$, $n = 336$)	2	MI-C	23.7	17.1	17.1	23.2	15.8	15.8	23.2	15.8	20.2	23.2	15.8	15.8
		MI-D	23.7	18.4	18.9	23.7	18.4	18.4	23.7	17.1	18.3	23.7	17.6	18.9
	4	MI-C	10.1	8.5	9.5	9.5	8.5	9.3	9.5	8.1	9.5	9.5	9.1	9.7
		MI-D	11.7	11.7	11.7	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1	10.1
	6	MI-C	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5
		MI-D	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5	6.5
Letter ($\mathcal{D} = 16$, $n = 15000$)	5	MI-C	27.5	26.1	26.1	29.7	24.6	24.6	24.6	24.6	24.6	27.5	26.1	26.1
		MI-D	34.5	31.6	33.9	33.9	30.1	30.1	29.9	29.9	25.0	34.5	31.6	32.5
	10	MI-C	14.7	13.9	14.1	13.9	13.6	13.7	13.7	12.9	13.1	13.9	13.4	13.4
		MI-D	15.2	14.1	14.1	13.9	13.9	13.9	13.9	13.2	13.6	13.9	13.6	13.6
	15	MI-C	12.4	12.4	12.4	12.4	12.4	12.4	12.4	12.1	12.4	12.4	12.4	12.4
		MI-D	12.4	12.4	12.4	12.4	12.4	12.4	12.4	12.4	12.4	12.4	12.4	12.4
Satimage ($\mathcal{D} = 36$, $n = 4435$)	5	MI-C	19.9	19.1	19.1	18.6	17.9	18.6	19.1	18.9	18.9	17.2	17.2	17.2
		MI-D	23.8	22.9	22.9	22.3	21.4	21.6	21.8	20.2	21.1	21.8	20.6	21.0
	10	MI-C	20.6	18.8	18.8	20.6	18.8	18.1	19.8	17.3	17.3	18.7	18.1	18.6
		MI-D	22.9	22.7	22.9	21.9	19.8	20.7	20.5	18.1	19.9	22.9	20.1	20.7
	15	MI-C	18.1	17.9	18.0	18.0	17.6	17.9	22.1	17.4	21.6	19.3	16.9	16.9
		MI-D	21.9	21.6	21.6	21.6	19.1	20.2	22.6	21.5	21.8	22.7	21.5	21.8
	20	MI-C	18.1	17.9	17.9	17.6	17.6	17.9	19.8	18.1	18.9	19.1	16.6	18.1
		MI-D	20.3	20.3	20.3	20.3	20.2	20.0	20.1	19.6	19.7	20.4	19.8	20.2
	25	MI-C	17.9	17.9	17.9	17.6	16.6	16.6	16.2	15.9	16.1	16.2	16.0	16.1
		MI-D	18.8	18.8	18.0	18.1	17.4	17.8	17.9	17.2	17.3	18.0	17.2	17.6
	30	MI-C	19.6	18.9	19.2	18.7	18.6	18.6	18.6	18.6	18.6	18.8	18.8	18.8
		MI-D	19.6	19.6	19.6	18.9	18.9	18.9	18.9	18.9	18.9	18.9	18.9	18.9
	35	MI-C	17.6	17.6	17.6	17.6	17.6	17.6	17.6	17.6	17.6	17.2	17.2	17.2
		MI-D	17.6	17.6	17.6	17.6	17.6	17.6	17.6	17.6	17.6	17.2	17.2	17.2
Isolet ($\mathcal{D} = 617$, $n = 7797$)	15	MI-C	16.3	11.9	12.1	16.3	11.5	12.1	15.7	11.3	12.0	15.1	11.4	12.1
		MI-D	23.7	23.7	22.3	23.7	19.1	22.3	21.4	18.5	18.9	22.0	18.8	18.8
	20	MI-C	18.7	14.0	14.7	18.9	13.1	15.7	18.3	12.7	13.3	18.4	12.8	13.9
		MI-D	24.3	22.8	22.8	24.3	19.6	21.0	22.3	19.1	19.6	22.5	19.3	20.5
	25	MI-C	15.7	14.5	14.5	12.9	9.6	11.7	12.1	8.8	11.6	12.5	9.1	11.9
		MI-D	21.1	19.2	21.1	21.1	17.1	18.3	20.0	12.3	19.5	19.6	14.1	19.1
	30	MI-C	11.7	8.2	11.2	11.7	8.2	10.6	10.9	8.0	10.4	11.4	8.3	10.5
		MI-D	17.3	14.5	14.5	17.3	14.5	16.9	14.3	11.7	12.8	15.0	11.8	12.9
	35	MI-C	11.2	7.6	11.5	11.5	6.9	10.6	10.9	6.8	10.3	11.2	6.8	10.7
		MI-D	14.7	9.2	11.2	14.7	9.2	11.3	13.2	8.3	10.7	13.3	9.0	10.9

These measures are applied to calculate both relevance and redundancy of the features. The values of α investigated are 0.2, 0.5, 0.8, 1.5, 2.0, 3.0, 4.0, and 5.0. The values close to 1.0 are excluded, either because the measures resemble mutual information for such values (I_α, $\mathcal{R}_\alpha$) or because they resemble another measure (M_1 and χ^1 equal VI). The performance of the proposed method is also compared with that of quick reduct algorithm, both in fuzzy (fuzzy-rough quick reduct) [10] and crisp (rough quick reduct) [18] approximation spaces.

To analyze the performance of proposed method, the experimentation is done on Iris, E-Coli, Wine, Letter, Ionosphere, Satimage, and Isolet data sets that are downloaded from http://www.ics.uci.edu/~mlearn. The major metrics for evaluating the performance of different algorithms are the proposed indexes, as well as some existing measures reported in Section 6. To compute the classification error of both K-NN rule and C4.5, the leave-one-out cross validation is performed on E-Coli, Wine, and Ionosphere data, while the training-testing is done on Letter and Satimage data.

7.1 Result on Iris Data

The parameters generated in the proposed feature selection method and the relevance of each feature are reported next for Iris data, as an example. The values of input parameters used are also presented here. The mutual information is chosen to calculate the relevance and redundancy of the features.

Number of samples (objects), $n = 150$
Number of dimensions (features), $\mathcal{D} = 4$
Value of weight parameter $\beta = 0.5$
Value of multiplicative parameter $\eta = 1.5$
Feature 1:
$\bar{c}_{\text{low}} = 0.2496$; $\bar{c}_{\text{medium}} = 0.4287$; $\bar{c}_{\text{high}} = 0.6333$
$\sigma_{\text{low}} = 0.3581$; $\sigma_{\text{medium}} = 0.5701$; $\sigma_{\text{high}} = 0.4093$
Feature 2:
$\bar{c}_{\text{low}} = 0.3138$; $\bar{c}_{\text{medium}} = 0.4392$; $\bar{c}_{\text{high}} = 0.5945$
$\sigma_{\text{low}} = 0.2508$; $\sigma_{\text{medium}} = 0.4157$; $\sigma_{\text{high}} = 0.3107$
Feature 3:
$\bar{c}_{\text{low}} = 0.1192$; $\bar{c}_{\text{medium}} = 0.4676$; $\bar{c}_{\text{high}} = 0.6811$
$\sigma_{\text{low}} = 0.6967$; $\sigma_{\text{medium}} = 0.8559$; $\sigma_{\text{high}} = 0.4269$
Feature 4:
$\bar{c}_{\text{low}} = 0.1146$; $\bar{c}_{\text{medium}} = 0.4578$; $\bar{c}_{\text{high}} = 0.6866$
$\sigma_{\text{low}} = 0.6864$; $\sigma_{\text{medium}} = 0.8725$; $\sigma_{\text{high}} = 0.4576$
Relevance of each feature:
Feature 1: 0.2669; Feature 2: 0.1488
Feature 3: 0.3793; Feature 4: 0.3739

TABLE 2
Performance on Satimage and Isolet Databases for Different Values of Weight Parameter β Considering $\eta = 1.5$

Data Set	Evaluation Criteria	Measure	Value of Weight Parameter β										
			0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Satimage ($d = 10$)	Classification Error (C4.5)	MI-C	19.8	19.8	19.8	19.1	18.5	17.3	17.3	17.3	17.3	17.3	17.3
		MI-D	20.5	20.5	20.5	20.5	19.7	18.1	18.1	18.8	19.9	19.9	19.9
		VI-C	26.6	26.6	26.6	22.9	17.4	17.4	17.4	17.4	16.8	16.8	16.8
		VI-D	29.6	29.6	29.6	25.1	18.1	18.1	18.1	18.1	23.7	23.7	25.1
	Class Separability	MI-C	0.393	0.393	0.393	0.393	0.384	0.366	0.366	0.366	0.366	0.366	0.366
		MI-D	0.597	0.597	0.597	0.597	0.593	0.467	0.467	0.467	0.467	0.464	0.464
		VI-C	0.417	0.417	0.397	0.397	0.366	0.366	0.366	0.366	0.366	0.366	0.366
		VI-D	0.643	0.643	0.643	0.617	0.495	0.467	0.495	0.495	0.560	0.560	0.560
	Entropy (E)	MI-C	0.828	0.828	0.828	0.828	0.828	0.824	0.824	0.824	0.824	0.827	0.827
		MI-D	0.833	0.833	0.833	0.833	0.830	0.832	0.832	0.830	0.830	0.830	0.830
		VI-C	0.817	0.817	0.811	0.802	0.802	0.802	0.802	0.809	0.809	0.809	0.809
		VI-D	0.832	0.832	0.832	0.832	0.830	0.830	0.829	0.829	0.830	0.827	0.830
	Representation Entropy (H_R)	MI-C	3.366	3.366	3.366	3.366	3.366	3.366	3.399	3.399	3.399	3.399	3.399
		MI-D	3.260	3.260	3.263	3.263	3.263	3.263	3.263	3.298	3.298	3.298	3.298
		VI-C	3.344	3.344	3.344	3.420	3.420	3.420	3.420	3.406	3.366	3.366	3.366
		VI-D	3.198	3.198	3.198	3.217	3.217	3.217	3.226	3.226	3.226	3.226	3.226
	$\mathbb{RELEV}$ Index	MI-C	0.427	0.427	0.427	0.427	0.427	0.427	0.425	0.425	0.425	0.425	0.425
		MI-D	0.338	0.338	0.336	0.334	0.334	0.334	0.334	0.330	0.330	0.330	0.330
		VI-C	0.439	0.439	0.439	0.438	0.434	0.434	0.434	0.434	0.431	0.431	0.430
		VI-D	0.343	0.343	0.343	0.343	0.340	0.340	0.340	0.339	0.339	0.339	0.339
	$\mathbb{REDUN}$ Index	MI-C	0.350	0.350	0.350	0.350	0.350	0.345	0.345	0.345	0.345	0.345	0.345
		MI-D	0.419	0.419	0.419	0.419	0.402	0.350	0.350	0.350	0.350	0.350	0.348
		VI-C	0.423	0.423	0.423	0.419	0.419	0.419	0.406	0.406	0.399	0.399	0.399
		VI-D	0.428	0.428	0.428	0.428	0.428	0.411	0.411	0.418	0.408	0.408	0.408
Isolet ($d = 25$)	Classification Error (C4.5)	MI-C	12.1	12.1	10.4	9.7	8.8	8.8	8.7	8.7	10.1	10.3	11.6
		MI-D	20.0	20.0	18.6	12.3	12.3	12.3	11.9	13.7	16.2	19.5	19.5
		VI-C	11.5	10.2	9.7	9.7	9.7	8.4	8.3	8.5	8.4	8.4	8.4
		VI-D	18.6	18.6	14.9	13.1	11.4	11.4	11.4	15.0	16.4	17.3	17.3
	Class Separability	MI-C	0.158	0.158	0.147	0.140	0.126	0.113	0.113	0.113	0.113	0.113	0.113
		MI-D	0.371	0.371	0.371	0.371	0.344	0.344	0.344	0.344	0.344	0.344	0.344
		VI-C	0.138	0.138	0.138	0.127	0.127	0.097	0.097	0.097	0.114	0.114	0.123
		VI-D	0.362	0.362	0.358	0.358	0.358	0.358	0.359	0.357	0.357	0.357	0.357
	Entropy (E)	MI-C	0.276	0.276	0.276	0.276	0.276	0.276	0.275	0.275	0.275	0.275	0.275
		MI-D	0.313	0.313	0.313	0.313	0.313	0.313	0.313	0.313	0.311	0.309	0.309
		VI-C	0.276	0.276	0.276	0.276	0.276	0.276	0.273	0.273	0.273	0.273	0.273
		VI-D	0.304	0.304	0.304	0.304	0.303	0.303	0.301	0.301	0.301	0.301	0.301
	Representation Entropy (H_R)	MI-C	4.619	4.619	4.619	4.619	4.629	4.629	4.629	4.629	4.629	4.629	4.629
		MI-D	4.402	4.402	4.402	4.407	4.407	4.407	4.410	4.410	4.410	4.410	4.410
		VI-C	4.637	4.641	4.641	4.644	4.647	4.647	4.647	4.630	4.630	4.630	4.612
		VI-D	4.441	4.441	4.441	4.441	4.445	4.445	4.445	4.445	4.439	4.439	4.439
	$\mathbb{RELEV}$ Index	MI-C	0.417	0.417	0.417	0.417	0.417	0.417	0.419	0.419	0.419	0.419	0.419
		MI-D	0.311	0.311	0.311	0.311	0.314	0.314	0.314	0.314	0.314	0.314	0.314
		VI-C	0.416	0.416	0.417	0.417	0.417	0.418	0.418	0.415	0.415	0.411	0.409
		VI-D	0.313	0.313	0.313	0.313	0.316	0.316	0.316	0.316	0.316	0.316	0.316
	$\mathbb{REDUN}$ Index	MI-C	0.427	0.427	0.427	0.427	0.425	0.425	0.425	0.425	0.425	0.425	0.425
		MI-D	0.496	0.496	0.496	0.490	0.490	0.490	0.490	0.490	0.490	0.490	0.490
		VI-C	0.427	0.427	0.427	0.424	0.424	0.424	0.424	0.424	0.424	0.425	0.425
		VI-D	0.462	0.462	0.462	0.462	0.461	0.461	0.461	0.458	0.458	0.458	0.458

In the proposed feature selection method, Feature 3 will be selected first as it has the highest relevance value. After selecting Feature 3, the redundancy and objective function of each feature are calculated that follow next.

Redundancy of each feature:
Feature 1: 0.1295; Feature 2: 0.0572; Feature 4: 0.1522
Value of objective function:
Feature 1: 0.2021; Feature 2: 0.1202; Feature 4: 0.2978

Based on the value of objective function, Feature 4 will be selected next as the second feature. The values of different quantitative indexes for these two features (Features 3 and 4) are reported next, along with that for whole feature sets.

Measures/Features	3 and 4	1 to 4
Classification error, C4.5	2.0%	2.0%
Class separability, S	0.0909	0.2343
Entropy, E	0.6904	0.7535
Representation Entropy, H_R	0.9973	0.8785
$\mathbb{RELEV}$ Index	0.5126	0.4407
$\mathbb{REDUN}$ Index	0.4149	0.4440

The results reported above establish the fact that the proposed method selects most significant features from the whole feature sets by maximizing the relevance and minimizing the redundancy of selected features.

7.2 Effectiveness of the Proposed Method

To better understand the effectiveness of the proposed method, extensive experimental results are reported in

TABLE 3
Performance on Satimage and Isolet Databases for Different Values of Multiplicative Parameter η Considering $\beta = 0.5$

Data Set	Evaluation Criteria	Measure	Value of Multiplicative Parameter η											
			0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
Satimg ($d = 10$)	Classification Error (C4.5)	MI-C	18.8	18.8	18.8	18.8	19.3	18.6	17.2	17.3	17.1	18.5	18.6	18.1
		MI-D	22.7	22.7	20.9	19.8	19.8	19.8	18.4	18.1	18.2	19.7	20.1	20.1
		VI-C	17.8	17.8	17.8	17.8	17.8	17.6	17.4	17.4	18.6	17.2	17.2	17.4
		VI-D	21.3	21.3	21.3	20.8	20.8	18.6	18.2	18.1	19.7	19.2	19.2	19.0
	Class Separability	MI-C	0.449	0.467	0.458	0.465	0.367	0.367	0.367	0.366	0.310	0.408	0.421	0.461
		MI-D	0.617	0.611	0.611	0.608	0.563	0.563	0.467	0.467	0.460	0.499	0.586	0.586
		VI-C	0.447	0.463	0.417	0.417	0.398	0.398	0.361	0.366	0.355	0.328	0.391	0.437
		VI-D	0.614	0.610	0.632	0.617	0.559	0.510	0.483	0.467	0.491	0.491	0.512	0.580
	Entropy (E)	MI-C	0.836	0.832	0.832	0.832	0.830	0.829	0.828	0.824	0.832	0.839	0.840	0.838
		MI-D	0.861	0.863	0.844	0.844	0.844	0.834	0.830	0.832	0.837	0.841	0.841	0.846
		VI-C	0.835	0.835	0.834	0.829	0.829	0.816	0.816	0.802	0.802	0.811	0.834	0.833
		VI-D	0.858	0.858	0.851	0.850	0.841	0.841	0.837	0.830	0.827	0.832	0.832	0.839
	Representation Entropy (H_R)	MI-C	3.284	3.284	3.284	3.284	3.284	3.282	3.313	3.366	3.295	3.295	3.299	3.299
		MI-D	3.217	3.217	3.246	3.248	3.248	3.263	3.263	3.263	3.263	3.254	3.259	3.259
		VI-C	3.321	3.321	3.299	3.325	3.327	3.327	3.420	3.420	3.420	3.421	3.417	3.417
		VI-D	3.208	3.208	3.208	3.212	3.219	3.214	3.214	3.217	3.217	3.217	3.211	3.211
	$\mathbb{RELEV}$ Index	MI-C	0.257	0.283	0.307	0.331	0.358	0.383	0.407	0.427	0.407	0.387	0.393	0.403
		MI-D	0.254	0.254	0.278	0.305	0.314	0.327	0.334	0.334	0.334	0.303	0.318	0.318
		VI-C	0.421	0.419	0.421	0.431	0.431	0.434	0.434	0.434	0.437	0.437	0.429	0.429
		VI-D	0.261	0.260	0.285	0.308	0.313	0.330	0.337	0.340	0.344	0.326	0.319	0.318
	$\mathbb{REDUN}$ Index	MI-C	0.369	0.393	0.419	0.443	0.369	0.392	0.412	0.345	0.408	0.409	0.424	0.437
		MI-D	0.417	0.423	0.428	0.441	0.397	0.378	0.365	0.350	0.389	0.414	0.451	0.455
		VI-C	0.451	0.437	0.434	0.426	0.425	0.422	0.417	0.419	0.424	0.437	0.440	0.448
		VI-D	0.460	0.449	0.440	0.429	0.428	0.422	0.416	0.411	0.408	0.426	0.443	0.459
Isolet ($d = 25$)	Classification Error (C4.5)	MI-C	14.5	14.5	12.7	9.6	12.4	11.5	9.3	8.8	9.2	8.9	9.1	9.1
		MI-D	19.2	19.2	18.4	17.1	17.0	13.8	12.6	12.3	12.3	13.9	13.9	14.1
		VI-C	12.4	12.3	10.9	9.4	9.4	8.8	8.7	8.4	8.3	8.5	9.1	9.0
		VI-D	18.7	18.7	17.4	16.3	16.1	13.9	12.0	11.4	11.5	11.4	12.8	13.7
	Class Separability	MI-C	0.259	0.241	0.268	0.255	0.218	0.189	0.147	0.113	0.126	0.169	0.179	0.208
		MI-D	0.411	0.403	0.395	0.374	0.368	0.351	0.350	0.344	0.341	0.363	0.398	0.418
		VI-C	0.153	0.158	0.131	0.118	0.113	0.105	0.099	0.097	0.097	0.102	0.134	0.162
		VI-D	0.411	0.417	0.404	0.396	0.388	0.364	0.361	0.358	0.357	0.391	0.399	0.412
	Entropy (E)	MI-C	0.286	0.286	0.287	0.284	0.284	0.281	0.279	0.276	0.276	0.278	0.283	0.289
		MI-D	0.329	0.328	0.329	0.322	0.322	0.318	0.314	0.313	0.316	0.317	0.320	0.324
		VI-C	0.286	0.286	0.286	0.283	0.279	0.279	0.278	0.276	0.276	0.277	0.281	0.285
		VI-D	0.317	0.316	0.316	0.314	0.315	0.308	0.308	0.303	0.304	0.311	0.312	0.317
	Representation Entropy (H_R)	MI-C	4.403	4.417	4.329	4.461	4.466	4.607	4.633	4.629	4.617	4.502	4.495	4.327
		MI-D	4.017	4.004	4.097	4.176	4.205	4.289	4.331	4.407	4.416	4.228	4.109	4.082
		VI-C	4.414	4.427	4.326	4.475	4.483	4.591	4.603	4.647	4.644	4.428	4.417	4.445
		VI-D	4.016	4.017	4.073	4.128	4.362	4.360	4.447	4.445	4.446	4.437	4.308	4.184
	$\mathbb{RELEV}$ Index	MI-C	0.403	0.399	0.397	0.391	0.404	0.403	0.415	0.417	0.417	0.411	0.410	0.394
		MI-D	0.296	0.288	0.292	0.306	0.311	0.308	0.315	0.314	0.312	0.299	0.278	0.263
		VI-C	0.404	0.402	0.394	0.396	0.401	0.407	0.416	0.418	0.415	0.412	0.411	0.405
		VI-D	0.296	0.289	0.295	0.299	0.307	0.311	0.312	0.316	0.315	0.311	0.298	0.291
	$\mathbb{REDUN}$ Index	MI-C	0.433	0.429	0.446	0.452	0.427	0.422	0.425	0.425	0.426	0.433	0.461	0.472
		MI-D	0.514	0.506	0.507	0.502	0.495	0.499	0.491	0.490	0.486	0.497	0.512	0.521
		VI-C	0.437	0.433	0.429	0.461	0.452	0.437	0.429	0.424	0.428	0.455	0.463	0.470
		VI-D	0.497	0.502	0.506	0.499	0.476	0.469	0.466	0.461	0.467	0.475	0.471	0.489

Table 1. Subsequent discussions analyze the results with respect to the classification error of C4.5.

Table 1 reports the classification error of C4.5 for mutual-information-based feature selection method both in fuzzy and crisp approximation spaces. Results are presented for different values of the number of selected features d, weight parameter β, and multiplicative parameter η. All the results reported here confirm that mutual-information-based feature selection method is more effective in fuzzy approximation spaces than in crisp approximation spaces with smaller number of features. The proposed feature selection method in fuzzy approximation spaces improves the classification accuracy of C4.5 significantly over its crisp counterpart, especially at smaller number of features. As the number of selected features d increases, the difference between fuzzy and crisp approximation spaces decreases. For a given data set with n samples and $\mathcal{D}$ features, the classification error of C4.5 remains unchanged for any combination of β and η when the number of selected features d approaches to $\mathcal{D}$. In case of E-Coli and Letter data sets, the error becomes almost same for $d = 6$ and 15 as the values of corresponding $\mathcal{D} = 7$ and 16, respectively. Similarly, for Satimage data set, the classification error remains almost same at $d = 35$ as the corresponding $\mathcal{D} = 36$. However, for feature selection, small feature set is of practical importance. Also, for a given data set and fixed d and η values, the classification error would be lower for nonzero β values. In other words, if the redundancy between the selected feature sets is taken into consideration, the performance of the proposed method would be better both in fuzzy and crisp approximation spaces.

7.3 Optimum Value of Weight Parameter β

The parameter β regulates the relative importance of the redundancy between the candidate feature and the already selected features with respect to the relevance with the

TABLE 4
Comparative Performance Analysis of Different f-Information Measures on Letter Database for $d = 6$

Type of f	Value of α	C4.5 Error		K-NN Error		Separability		Entropy E		Entropy H_R		$\mathbb{RELEV}$ Index		$\mathbb{REDUN}$ Index	
		fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp
I_α	0.2	21.8	25.2	11.6	16.2	1.159	1.608	0.889	0.911	3.208	3.082	0.170	0.164	0.237	0.301
	0.5	16.3	19.1	7.8	15.2	1.087	1.108	0.881	0.898	3.316	3.264	0.231	0.219	0.207	0.283
	0.8	21.8	25.2	11.6	16.2	1.159	1.608	0.889	0.911	3.208	3.082	0.170	0.164	0.237	0.301
	1.0	23.2	26.7	16.9	22.8	1.345	1.562	0.892	0.914	3.172	3.005	0.137	0.118	0.299	0.316
	1.5	21.8	25.2	11.6	16.2	1.159	1.608	0.889	0.911	3.208	3.082	0.170	0.164	0.237	0.301
	2.0	23.2	26.7	16.9	22.8	1.345	1.562	0.892	0.914	3.172	3.005	0.137	0.118	0.299	0.316
	3.0	23.2	26.7	16.9	22.8	1.345	1.562	0.892	0.914	3.172	3.005	0.137	0.118	0.299	0.316
M_α	0.2	21.8	25.2	11.6	16.2	1.159	1.608	0.889	0.911	3.208	3.082	0.170	0.164	0.237	0.301
	0.5	16.3	19.1	7.8	15.2	1.087	1.108	0.881	0.898	3.316	3.264	0.231	0.219	0.207	0.283
	0.8	21.8	25.2	11.6	16.2	1.159	1.608	0.889	0.911	3.208	3.082	0.170	0.164	0.237	0.301
	1.0	17.5	21.3	8.7	17.2	1.055	1.114	0.874	0.903	3.469	3.187	0.240	0.213	0.193	0.294
χ^α	1.5	21.8	25.2	11.6	16.2	1.159	1.608	0.889	0.911	3.208	3.082	0.170	0.164	0.237	0.301
	2.0	23.2	26.7	16.9	22.8	1.345	1.562	0.892	0.914	3.172	3.005	0.137	0.118	0.299	0.316
	3.0	23.2	26.7	16.9	22.8	1.345	1.562	0.892	0.914	3.172	3.005	0.137	0.118	0.299	0.316
$\mathcal{R}_\alpha$	0.2	21.8	25.2	11.6	16.2	1.159	1.608	0.889	0.911	3.208	3.082	0.170	0.164	0.237	0.301
	0.5	21.8	25.2	11.6	16.2	1.159	1.608	0.889	0.911	3.208	3.082	0.170	0.164	0.237	0.301
	0.8	21.8	25.2	11.6	16.2	1.159	1.608	0.889	0.911	3.208	3.082	0.170	0.164	0.237	0.301
	1.0	23.2	26.7	16.9	22.8	1.345	1.562	0.892	0.914	3.172	3.005	0.137	0.118	0.299	0.316
	1.5	21.8	25.2	11.6	16.2	1.159	1.608	0.889	0.911	3.208	3.082	0.170	0.164	0.237	0.301
	2.0	23.2	26.7	16.9	22.8	1.345	1.562	0.892	0.914	3.172	3.005	0.137	0.118	0.299	0.316
	3.0	23.2	26.7	16.9	22.8	1.345	1.562	0.892	0.914	3.172	3.005	0.137	0.118	0.299	0.316

TABLE 5
Comparative Performance Analysis of Different f-Information Measures on Satimage Database for $d = 10$

Type of f	Value of α	C4.5 Error		K-NN Error		Separability		Entropy E		Entropy H_R		$\mathbb{RELEV}$ Index		$\mathbb{REDUN}$ Index	
		fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp
I_α	0.2	18.0	17.9	17.2	18.0	0.435	0.465	0.828	0.827	3.298	3.257	0.419	0.341	0.349	0.349
	0.5	18.6	18.1	17.2	18.8	0.369	0.467	0.829	0.832	3.282	3.248	0.426	0.334	0.345	0.350
	0.8	17.3	18.1	17.3	18.8	0.367	0.467	0.828	0.832	3.263	3.263	0.427	0.334	0.346	0.350
	1.0	17.3	18.1	17.3	18.8	0.366	0.467	0.824	0.832	3.366	3.263	0.427	0.334	0.345	0.350
	1.5	17.2	18.6	14.8	16.1	0.360	0.478	0.829	0.829	3.364	3.254	0.427	0.329	0.347	0.357
	2.0	17.2	18.6	14.8	16.1	0.361	0.478	0.801	0.829	3.364	3.254	0.428	0.329	0.346	0.357
	3.0	17.1	18.0	17.1	18.0	0.361	0.461	0.827	0.827	3.366	3.260	0.428	0.330	0.341	0.407
	4.0	16.6	18.4	13.6	16.0	0.359	0.473	0.801	0.827	3.458	3.248	0.437	0.324	0.336	0.372
	5.0	26.1	18.4	19.5	16.1	0.447	0.473	0.831	0.827	3.197	3.248	0.411	0.324	0.425	0.372
M_α	0.2	19.8	17.7	16.6	18.1	0.418	0.462	0.824	0.828	3.117	3.351	0.422	0.339	0.379	0.413
	0.5	18.6	18.1	17.2	18.8	0.369	0.467	0.829	0.832	3.282	3.248	0.426	0.334	0.345	0.350
	0.8	17.4	18.1	16.6	18.1	0.369	0.467	0.807	0.830	3.289	3.217	0.429	0.340	0.416	0.411
	1.0	17.4	18.1	16.6	18.1	0.366	0.467	0.802	0.830	3.420	3.217	0.434	0.340	0.419	0.411
χ^α	1.5	17.4	18.1	15.1	18.8	0.364	0.467	0.806	0.830	3.282	3.278	0.427	0.337	0.346	0.409
	2.0	17.2	18.6	14.8	16.1	0.361	0.478	0.801	0.829	3.364	3.254	0.428	0.329	0.346	0.357
	3.0	16.6	18.6	13.6	16.1	0.360	0.478	0.797	0.829	3.455	3.254	0.434	0.329	0.389	0.357
	4.0	26.1	18.6	19.5	16.1	0.441	0.478	0.816	0.829	3.197	3.254	0.408	0.329	0.312	0.357
	5.0	30.9	18.6	23.7	16.1	0.486	0.478	0.823	0.829	3.068	3.254	0.401	0.329	0.317	0.357
$\mathcal{R}_\alpha$	0.2	18.0	17.9	17.2	18.0	0.435	0.465	0.828	0.827	3.298	3.298	0.427	0.341	0.349	0.349
	0.5	18.6	18.1	17.2	18.8	0.369	0.467	0.829	0.832	3.282	3.261	0.426	0.334	0.345	0.350
	0.8	17.3	18.1	17.3	18.8	0.367	0.467	0.828	0.832	3.263	3.263	0.427	0.334	0.346	0.350
	1.0	17.3	18.1	17.3	18.8	0.366	0.467	0.824	0.832	3.366	3.263	0.427	0.334	0.345	0.350
	1.5	17.2	20.6	14.1	19.7	0.360	0.491	0.829	0.834	3.364	3.118	0.428	0.326	0.347	0.351
	2.0	17.2	20.6	14.1	23.6	0.361	0.491	0.826	0.834	3.364	3.118	0.428	0.326	0.347	0.351
	3.0	16.6	18.6	13.9	16.1	0.359	0.478	0.801	0.829	3.457	3.254	0.435	0.329	0.339	0.357
	4.0	16.4	18.6	13.6	16.1	0.355	0.478	0.794	0.831	3.478	3.254	0.441	0.329	0.338	0.357
	5.0	27.2	18.6	21.5	16.1	0.497	0.478	0.828	0.829	3.206	3.254	0.408	0.329	0.430	0.357

output class. If β is zero, only the relevance with the output class is considered for each feature. If β increases, this measure is discounted by a quantity proportional to the total redundancy with respect to the already selected features. The value of β larger than zero is crucial in order to obtain good results. If the redundancy between features is not taken into account, selecting the features with the highest relevance with respect to the output class tends to produce a set of redundant features that may leave out useful complementary information.

Table 2 presents the performance of proposed method using both V and mutual information for different values of β. The results and subsequent discussions are presented in this table with respect to various proposed and existing quantitative indexes for both fuzzy and crisp approximation spaces. In Table 2, it is seen that as the value of β increases, the values of $\mathbb{RELEV}$ index and representative entropy H_R increase, whereas the classification error of C4.5, the values of $\mathbb{REDUN}$ index, class separability S, and entropy E decrease. The V and mutual information achieve their best performance for $0.5 \leq \beta < 1$ with respect

TABLE 6
Comparative Performance Analysis of Different Methods Using Proposed and Existing Feature Evaluation Indexes

Data Sets	Method/ Measure	K-NN Error		C4.5 Error		Separability		$\mathbb{RELEV}$ Index		Entropy E		Entropy H_R		$\mathbb{REDUN}$ Index	
		fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp
Wine	MI	7.3	7.3	7.9	7.9	0.181	0.181	0.443	0.443	0.743	0.743	0.997	0.997	0.413	0.443
	$I_{4.0}$	4.1	5.8	6.7	7.9	0.138	0.181	0.442	0.443	0.756	0.743	0.985	0.997	0.406	0.443
	$M_{0.8}$/VI	3.9	5.8	2.8	4.5	0.108	0.154	0.445	0.434	0.741	0.745	0.998	0.997	0.417	0.430
	$\chi^{1.5}$	3.9	5.8	2.8	4.5	0.108	0.154	0.445	0.434	0.741	0.745	0.998	0.997	0.417	0.430
	$\mathcal{R}_{4.0}$	4.1	5.8	6.7	7.9	0.138	0.181	0.442	0.443	0.756	0.743	0.985	0.997	0.406	0.443
	QR	3.9	7.0	2.8	9.2	0.108	0.196	0.445	0.442	0.741	0.752	0.998	0.998	0.417	0.459
Letter	MI	16.9	22.8	23.2	26.7	1.345	1.562	0.137	0.118	0.892	0.914	3.172	3.005	0.299	0.316
	$I_{0.5}/M_{0.5}$	7.8	15.2	16.3	19.1	1.087	1.108	0.231	0.219	0.881	0.898	3.316	3.264	0.207	0.283
	VI	8.7	17.2	17.5	21.3	1.055	1.114	0.240	0.213	0.874	0.903	3.469	3.187	0.193	0.294
	$\mathcal{R}_{1.5}$	11.6	16.2	21.8	25.2	1.159	1.608	0.170	0.164	0.889	0.911	3.208	3.082	0.237	0.301
	QR	21.8	29.5	38.3	39.7	8.809	9.627	0.285	0.271	0.891	0.906	3.314	3.001	0.330	0.385
Ionsph	MI	2.6	4.9	2.6	4.8	1.655	3.283	0.393	0.387	0.765	0.765	3.298	3.293	0.374	0.375
	$I_{1.5}/M_{0.8}$	3.7	4.8	4.6	4.8	2.297	3.283	0.404	0.387	0.765	0.765	3.298	3.293	0.391	0.375
	VI	2.6	4.8	4.0	4.8	2.681	3.283	0.406	0.387	0.762	0.765	3.295	3.293	0.387	0.375
	$\chi^{1.5}/\mathcal{R}_{1.5}$	3.7	4.8	4.6	4.8	2.297	3.283	0.404	0.387	0.765	0.765	3.298	3.293	0.391	0.375
	QR	6.7	9.2	8.5	11.7	5.877	9.361	0.262	0.257	0.754	0.755	3.299	3.291	0.513	0.558
Satimg	MI	17.3	18.8	17.3	18.1	0.366	0.467	0.427	0.334	0.824	0.832	3.366	3.263	0.345	0.350
	$I_{4.0}$	13.6	16.0	16.6	18.4	0.359	0.473	0.437	0.324	0.801	0.827	3.458	3.248	0.336	0.372
	$M_{0.8}$	16.6	18.1	17.4	18.1	0.369	0.467	0.429	0.340	0.807	0.830	3.289	3.217	0.416	0.411
	VI	16.6	18.1	17.4	18.1	0.366	0.467	0.434	0.340	0.802	0.830	3.420	3.217	0.419	0.411
	$\chi^{3.0}$	13.6	16.1	16.6	18.6	0.360	0.478	0.434	0.329	0.797	0.829	3.455	3.254	0.389	0.357
	$\mathcal{R}_{4.0}$	13.6	16.1	16.4	18.6	0.355	0.478	0.441	0.329	0.794	0.831	3.478	3.254	0.338	0.357
	QR	19.2	24.8	21.6	24.8	0.892	0.996	0.458	0.357	0.795	0.834	3.118	3.006	0.513	0.529
Isolet	MI	6.1	9.3	8.8	12.3	0.113	0.344	0.417	0.314	0.276	0.313	4.629	4.407	0.425	0.490
	$I_{4.0}$	5.8	9.9	8.3	11.2	0.089	0.353	0.401	0.313	0.276	0.303	4.627	4.417	0.412	0.455
	$M_{0.8}$	7.8	9.3	9.5	11.4	0.128	0.389	0.417	0.316	0.276	0.303	4.641	4.445	0.426	0.463
	VI	5.8	10.7	8.4	11.4	0.097	0.358	0.418	0.316	0.276	0.303	4.647	4.445	0.424	0.461
	$\chi^{3.0}$	6.1	9.5	8.6	11.9	0.088	0.351	0.404	0.315	0.279	0.308	4.646	4.449	0.413	0.459
	$\mathcal{R}_{4.0}$	5.8	9.9	8.2	13.1	0.083	0.419	0.425	0.311	0.276	0.329	4.642	4.327	0.413	0.492
	QR	9.5	15.2	12.8	15.2	1.362	1.594	0.449	0.361	0.278	0.341	4.517	4.211	0.507	0.539

TABLE 7
Comparative Execution Time (in Millisecond) Analysis of Different Methods

Method / Measure	Wine ($d = 2$)		Letter ($d = 6$)		Ionosphere ($d = 10$)		Satimage ($d = 10$)		Isolet ($d = 25$)	
	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp	fuzzy	crisp
MI	8	7	2758	2699	163	144	2384	2273	143467	141973
I_α	7	7	2687	2685	165	147	2407	2239	143460	142157
M_α	8	8	2706	2776	168	153	2395	2192	143481	141996
VI	8	7	2694	2782	162	142	2383	2244	143459	142018
χ^α	8	8	2688	2695	167	146	2393	2248	143478	141980
$\mathcal{R}_\alpha$	7	8	2691	2795	167	150	2388	2267	143501	141769
QR	7	9	24568	19473	2483	2107	39179	37982	71412733	70488913

to all these quantitative indexes. In other words, the best performance of V and mutual information is achieved when the relevance of each feature is discounted by at least 50 percent of total redundancy with respect to already selected features.

7.4 Optimum Value of Multiplicative Parameter η

The η is a multiplicative parameter controlling the extent of overlapping between the fuzzy sets 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 π functions increases.

Table 3 represents the performance of the proposed method in terms of various quantitative indexes for different values of η. Results are presented for different data sets considering the information measure as both mutual information and V information. It is seen that in case of both mutual information and V-information, the proposed method achieves consistently better performance for $1.1 < \eta < 1.7$. In fact, very large or very small amounts of overlapping among the three fuzzy sets of the input feature are found to be undesirable.

7.5 Performance of Different f-Information

Furthermore, extensive experiments are done to evaluate the performance of different f information measures, both in fuzzy and crisp approximation spaces. Tables 4 and 5 report the results for different values of α considering $\beta = 0.5$ and $\eta = 1.5$. For each data set, the value of d (number of selected features) is chosen through extensive experimentation in such a way that the classification error of both C4.5 and K-NN becomes almost equal to that of original feature set.

From the results reported in Tables 4 and 5, it is seen that most of the f-information measures achieve consistently better performance than mutual information (= $I_{1.0}$- or

$\mathcal{R}_{1.0}$-information) for different values of α, both in fuzzy and crisp approximation spaces. Some f-information measures are shown to perform poorly on all aspects for certain values of α. The majority of measures produces results similar to those of mutual information. An important finding, however, is that several measures, although slightly more difficult to optimize, can potentially yield significantly better results than mutual information. For Satimage data, V- or $M_{1.0}$-, I_α- and $\mathcal{R}_\alpha$-information for $0.8 \leq \alpha \leq 4.0$, and χ^α-information for $\alpha = 2.0$ and 3.0 perform better than mutual information in fuzzy approximation spaces, while for Letter data, $I_{0.5}$-, $M_{0.5}$-, and V-information yield the best result with respect to most of the indexes and other measures are comparable to mutual information. However, the lowest value of $\mathbb{REDUN}$ index for Satimage data is achieved using $\chi^{4.0}$- and $\chi^{5.0}$-information.

7.6 Performance of Different Algorithms

Table 6 compares the best performance of different f-information that is used in the proposed feature selection method. The results are presented based on the minimum classification error of both C4.5 and K-NN. The values of β and η are considered as 0.5 and 1.5, respectively. The best performance of quick reduct (QR) algorithm, both in fuzzy [10] and crisp [18] approximation spaces, is also provided for the sake of comparison. It is seen that the f-information in fuzzy approximation spaces is more effective than that in crisp approximation spaces. The f-information-measure-based proposed feature selection method selects a set of features having the lowest classification error of both C4.5 and K-NN, class separability, entropy, and $\mathbb{REDUN}$ index values and the highest representation entropy and $\mathbb{RELEV}$ index values for all the cases. Also, several f-information measures, although slightly more difficult to optimize, can potentially yield significantly better results than mutual information, both in fuzzy and crisp approximation spaces. Moreover, the f-information-based proposed method outperforms quick reduct algorithm, both in fuzzy and crisp approximation spaces. However, quick reduct algorithm achieves the best $\mathbb{RELEV}$ index value for all data sets as it selects only relevant features of a data set without considering the redundancy among them. The better performance of the proposed method using f-information is achieved due to the fact that the fuzzy equivalence partition matrix provides an efficient way to calculate different f-information measures on fuzzy approximation spaces. In effect, a reduced set of features having maximum relevance and minimum redundancy is being obtained using the proposed method. Finally, Table 7 reports the execution time of different algorithms. The significantly lesser time of the proposed algorithm is achieved due to its low computational complexity.

8 Conclusion

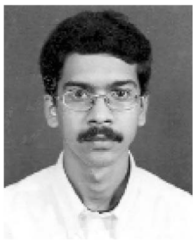
The problem of feature selection is highly important, particularly given the explosive growth of available information. In this paper, a novel feature selection method is presented based on fuzzy-rough sets. Using the concept of f-information measures on fuzzy approximation spaces, an efficient algorithm is introduced for finding nonredundant and relevant features of real-valued data sets. This formulation is geared toward maximizing the utility of rough sets, fuzzy sets, and information measures with respect to knowledge discovery tasks. Several quantitative indexes are defined based on fuzzy-rough sets to evaluate the performance of the proposed feature selection method on fuzzy approximation spaces for real-life data sets. Finally, the effectiveness of the proposed method is presented, along with a comparison with other related algorithms, on a set of real-life data.

Acknowledgments

Sankar K. Pal is a J.C. Bose Fellow of the Government of India.

References

[1] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification and Scene Analysis.* John Wiley & Sons, 1999.
[2] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach.* Prentice Hall, 1982.
[3] M. Dash and H. Liu, "Consistency Based Search in Feature Selection," *Artificial Intelligence,* vol. 151, nos. 1/2, pp. 155-176, 2003.
[4] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Trans. Neural Network,* vol. 5, no. 4, pp. 537-550, 1994.
[5] J.R. Quinlan, *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993.
[6] I. Vajda, *Theory of Statistical Inference and Information.* Kluwer Academic, 1989.
[7] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever, "f-Information Measures in Medical Image Registration," *IEEE Trans. Medical Imaging,* vol. 23, no. 12, pp. 1508-1516, Dec. 2004.
[8] P. Maji, "f-Information Measures for Efficient Selection of Discriminative Genes from Microarray Data," *IEEE Trans. Biomedical Eng.,* vol. 56, no. 4, pp. 1-7, Apr. 2009.
[9] Z. Pawlak, *Rough Sets, Theoretical Aspects of Reasoning About Data.* Kluwer, 1991.
[10] R. Jensen and Q. Shen, "Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approach," *IEEE Trans. Knowledge and Data Eng.,* vol. 16, no. 12, pp. 1457-1471, Dec. 2004.
[11] P. Maji and S.K. Pal, "Rough-Fuzzy C-Medoids Algorithm and Selection of Bio-Basis for Amino Acid Sequence Analysis," *IEEE Trans. Knowledge and Data Eng.,* vol. 19, no. 6, pp. 859-872, June 2007.
[12] D. Dubois and H. Prade, "Rough Fuzzy Sets and Fuzzy Rough Sets," *Int'l J. General Systems,* vol. 17, pp. 191-209, 1990.
[13] P. Maji and S.K. Pal, "Rough Set Based Generalized Fuzzy C-Means Algorithm and Quantitative Indices," *IEEE Trans. System, Man and Cybernetics, Part B: Cybernetics,* vol. 37, no. 6, pp. 1529-1540, Dec. 2007.
[14] Q. Hu, D. Yu, Z. Xie, and J. Liu, "Fuzzy Probabilistic Approximation Spaces and Their Information Measures," *IEEE Trans. Fuzzy Systems,* vol. 14, no. 2, pp. 191-201, 2007.
[15] C. Shannon and W. Weaver, *The Mathematical Theory of Communication.* Univ. of Illinois Press, 1964.
[16] S.K. Pal and S. Mitra, *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing.* Wiley, 1999.
[17] M. Dash and H. Liu, "Unsupervised Feature Selection," *Proc. Pacific Asia Conf. Knowledge Discovery and Data Mining,* pp. 110-121, 2000.
[18] A. Chouchoulas and Q. Shen, "Rough Set-Aided Keyword Reduction for Text Categorisation," *Applied Artificial Intelligence,* vol. 15, no. 9, pp. 843-873, 2001.



Pradipta Maji received the BSc (Hons) degree in physics, the MSc degree in electronics science, and the PhD degree in the area of computer science from Jadavpur University, India, in 1998, 2000, and 2005, respectively. Currently, he is an assistant professor in the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. His research interests include pattern recognition, computational biology and bioinformatics, medical image processing, cellular automata, soft computing, and so forth. He has published around 60 papers in international journals and conferences. He has received the 2006 Best Paper Award of the International Conference on Visual Information Engineering from The Institution of Engineering and Technology, United Kingdom, 2008 Microsoft Young Faculty Award from Microsoft Research Lab. India Pvt, and the 2009 Young Scientist Award from the National Academy of Sciences, India, and has been selected as the 2009 Associate of the Indian Academy of Sciences. He is also a reviewer of many international journals.

Sankar K. Pal received the PhD degree in radio physics and electronics from the University of Calcutta in 1979, and another PhD degree in electrical engineering along with DIC from Imperial College, University of London, in 1982. He is the director and a distinguished scientist of the Indian Statistical Institute. He founded the Machine Intelligence Unit and the Center for Soft Computing Research: A National Facility in the Institute in Calcutta. He worked at the University of California, Berkeley, and the University of Maryland, College Park, in 1986-1987; the NASA Johnson Space Center, Houston, Texas, in 1990-1992 and 1994; and the US Naval Research Laboratory, Washington, District of Columbia, in 2004. Since 1997, he has been serving as a distinguished visitor of the IEEE Computer Society for the Asia-Pacific Region, and held several visiting positions in Hong Kong and Australian universities. He is a fellow of the IEEE, the Academy of Sciences for the Developing World (TWAS), Italy, International Association for Pattern Recognition, International Association of Fuzzy Systems, and all the four National Academies for Science/Engineering in India. He is a coauthor of 14 books and more than 300 research publications in the areas of Pattern Recognition and Machine Learning, Image Processing, Data Mining and Web Intelligence, Soft Computing, and Bioinformatics. He has received the 1990 S.S. Bhatnagar Prize and many prestigious awards in India and abroad including the 1999 G.D. Birla Award, 1998 Om Bhasin Award, 1993 Jawaharlal Nehru Fellowship, 2000 Khwarizmi International Award from the Islamic Republic of Iran, 2000-2001 FICCI Award, 1993 Vikram Sarabhai Research Award, 1993 NASA Tech Brief Award, 1994 IEEE Transactions on Neural Networks Outstanding Paper Award, 1995 NASA Patent Application Award, 1997 IETE-R.L. Wadhwa Gold Medal, the 2001 INSA-S.H. Zaheer Medal, 2005-06 ISC-P.C. Mahalanobis Birth Centenary Award (Gold Medal) for Lifetime Achievement, 2007 J.C. Bose Fellowship of the Government of India, and 2008 Vigyan Ratna Award from Science & Culture Organization, West Bengal. He is an associate editor and the editor-in-chief of many international journals.