Pattern Recognition 45 (2012) 2690–2707

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Class-dependent rough-fuzzy granular space, dispersion index and classification

Sankar K. Pal [a], Saroj K. Meher [b,*], Soumitra Dutta [c]

[a] Center for Soft Computing Research, Indian Statistical Institute, Kolkata 700108, India
[b] Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore 560059, India
[c] INSEAD, Blvd de Constance, Fontainebleau 77305, France

ARTICLE INFO

Article history:
Received 16 March 2010
Received in revised form
16 November 2011
Accepted 24 December 2011
Available online 25 January 2012

Keywords:
Fuzzy information granulation
Rough neighborhood sets
Rough-fuzzy granular computing
Soft computing
Pattern recognition
Remote sensing

ABSTRACT

A new rough-fuzzy model for pattern classification based on granular computing is described in the present article. In this model, we propose the formulation of class-dependent granules in fuzzy environment. Fuzzy membership functions are used to represent the feature-wise belonging to different classes, thereby producing fuzzy granulation of the feature space. The fuzzy granules thus generated possess better class discriminatory information that is useful in pattern classification with overlapping classes. Neighborhood rough sets are used in the selection of a subset of granulated features that explore the local/contextual information from neighbor granules. The model thus explores mutually the advantages of class-dependent fuzzy granulation and neighborhood rough set. The superiority of the proposed model to other similar methods is established with seven completely labeled data sets, including a synthetic remote sensing image, and two partially labeled real remote sensing images collected from satellites. Various performance measures, including a new method of dispersion estimation, are used for comparative analysis. The new measure called "dispersion score" quantifies the nature of distribution of the classified patterns among different classes so that lower is the dispersion, better is the classifier. The proposed model learns well even with a lower percentage of training set that makes the system fast. The model is seen to have lowest dispersion measure (i.e., misclassified patterns are confined to minimum number of classes) compared to others; thereby reflecting well the overlapping characteristics of a class with others, and providing a strong clue for the class-wise performance improvement with available higher-level information. The statistical significance of the proposed model is also supported by the χ^2 test.

1. Introduction

Granular computing (GrC) refers to that where computation and operations are performed on information granules (clumps of similar objects or points). GrC has been changed rapidly from a label to conceptual and computational paradigm of study that deals with information and knowledge processing. Many researchers [1,2] have used GrC models to build efficient computational algorithms that can handle huge amounts of data, information and knowledge. These models mainly deal with the efficiency, effectiveness and robustness of using granules, such as classes, clusters, subsets, groups and intervals in problem solving [3].

GrC can be studied based on its notions of representation and process. However, the main task to be focussed is to construct and describe information granules, a process called information granulation [4–6] on which GrC is oriented. Specifically, granulation is governed by the principles according to which the models should exploit the tolerance for imprecision and employ the coarsest level of granulation, which are consistent with the allowable level of imprecision. Modes of information granulation in which the granules are crisp, play important roles in a wide variety of approaches and techniques. Although crisp information granulation has wide range of applications, it has a major blind spot [7]. More particularly it fails to reflect most of the processes of human reasoning and concept formation, where the granules are more appropriately fuzzy rather than crisp [8,9]. The fuzziness in granules and their values is the characteristic of the ways in which human concepts are formed, organized and manipulated. In fact fuzzy information granulation does not refer to single fuzzy granule rather it is about a collection of fuzzy granules which result from granulating a crisp or fuzzy object. Depending on the problems and whether the granules are fuzzy or crisp, one may have operations like granular fuzzy computing or fuzzy granular computing [10].

* Corresponding author. Tel.: +91 80 28483002; fax: +91 80 28484265.
E-mail address: saroj.meher@gmail.com (S.K. Meher).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.12.027

In the recent past many research works have been carried out in the construction of fuzzy granules. The process of fuzzy granulation involves the basic idea of generating a family of fuzzy granules from numerical features and transform them into fuzzy linguistic variables. These variables thus keep the semantics of the data and are easy to understand. Fuzzy information granulation has come up with an important concept in fuzzy set theories, rough set theories and the combination of both in recent years [5,7,10,11]. In a hybrid approach, Banerjee et al. [12] described the granulation of information with the combination of rough, fuzzy and neural networks. Here the authors proposed a new scheme of knowledge encoding in a fuzzy multilayer perceptron using rough set-theoretic concepts, and applied successfully in the classification of speech and synthetic data. Further, Pal and Mitra in [13] discussed the rough-fuzzy hybridization method of information granulation scheme for case generation. In this study, fuzzy set theory is used for linguistic representation of patterns, thereby producing a fuzzy granulation of the feature space and rough set theory is used to obtain dependency rules which model informative regions in the granulated feature space. Superiority of this algorithm in terms of classification accuracy and case generation and retrieval times is demonstrated on some real-life data sets. In general, the process of fuzzy granulation can be broadly categorized as class-dependent (CD) and class-independent (CI). Fuzzy sets are used in both cases for linguistic representation of patterns and generation of fuzzy granulation of the feature space. With CI granulation each feature of the pattern is described with the membership values corresponding to the overlapping partitions of the linguistic properties *low, medium* and *high* [14]. These overlapping functions along each axis generate the fuzzy granulated feature space in n-dimension and the granulated space contains 3^n granules. The degree of belongingness of a pattern to a granule is determined by the corresponding membership function. However, this process of granulation does not take care of the class belonging information of features to different classes. This may lead to a degradation of performance in pattern classification, particularly for data sets with highly overlapping classes. On the other hand, in CD granulation, each feature explores its class belonging information to different classes. In this process, features are described by the fuzzy sets equal to the number of classes, and individual class information is restored by the generated fuzzy granules.

Rough set theory, as proposed by Pawlak [5] (henceforth it will be abbreviated as PaRS), has been proven to be an effective tool for feature selection, knowledge discovery and rule extraction from categorical data [15]. The theory enables the discovery of data dependencies and performs the reduction/selection of features contained in a data set using the data alone, requiring no additional information. PaRS can be used as an effective tool to deal with both vagueness and uncertainty in data sets and to perform granular computation. PaRS based feature selection not only retains the representational power of the data, but also maintains its minimum redundancy [15]. However for the numerical data, PaRS theory can be used with the discretisation of data that results in the loss of information and introduction of noise. To deal with this, neighborhood rough set (NRS) [16,17] is found to be suitable that can deal with both numerical and categorical data sets without discretisation. The advantage of NRS is that it facilitates to gather the possible local information through neighbor granules that is useful for a better discrimination of patterns, particularly in class overlapping environment. Many attempts have been made in the use of NRS for information granulation and feature reduction [18,19].

1.1. Motivation and the proposed solution

As discussed earlier, fuzzy granulation of information has the inherent advantages of improved human-like reasoning than crisp granulation, and CD fuzzy granulation further enhances its analysis capability compared to CI fuzzy granulation. This motivated us to explore the CD form of fuzzy granulation model for pattern classification. Further, the suitability of NRS based feature selection in dealing with numerical data, compared to Pawlak's rough set has encouraged to explore NRS based method for pattern classification. Although, individually the granulation and feature selection methods have their own advantages, the hybridization of both the methods with a cost of insignificant computational complexities has leaded to a domain that aims to cumulate the individual advantages. In recent past, many research attempts have described the benefits of hybridizing techniques and successfully demonstrated their superiority over individual methods for pattern classification.

This encouraged us to build a hybridizing rough-fuzzy granular space using fuzzy CD granulation and NRS based feature selection. The model provides a synergistic integration of the merits of both fuzzy CD granulation and the theory of NRS, and the resulting output can be used as an input to any classifier. To demonstrate the effectiveness of the proposed rough-fuzzy granular space based model compared to other similar methods, we have used here different classifiers, such as k-nearest neighbor (k-NN) ($k=1$, 2 and 3) classifier, maximum likelihood (ML) classifier [20] and multi-layered perceptron (MLP) [21]. However, other classifiers may also be used. We have demonstrated the potential of the proposed model with seven completely labeled data sets, including a synthetic multispectral remote sensing image, and two partially labeled real multispectral remote sensing images. For multispectral images we have used the spectral (band) values as features. Various performance measures such as percentage of overall classification accuracy, Precision, Recall [22], kappa coefficient [23] and computation time are considered for completely labeled data sets. In this context, a new index called 'dispersion score' reflecting a different interpretation of the confusion matrix, is defined to measure the class-wise classifier performance. The dispersion measure quantifies the nature of distribution of the classified patterns among different classes. In addition to these measures, we have performed the statistical significance test using χ^2 for supporting the superiority of the proposed model. For partially labeled data sets, on the other hand, β index [24] and Davies–Bouldin (DB) index [25] are computed to validate the superiority of the proposed model to others.

The novelty of the present work lies with the following. First, based on the class dependency knowledge, fuzzy granulated features are generated. Second, the neighborhood rough sets are applied on these fuzzy granular feature sets for computing the approximate reducts that select a subset of features. Finally, a different interpretation of confusion matrix is described by defining a new measure of dispersion which not only provides an index of class-wise performance, but also enables one to correct some of the misclassified patterns with available higher-level analysis. The experimental results show that the proposed model provides improved classification accuracy in terms of the aforesaid quantitative measures, even with a smaller training set.

2. Proposed model for pattern classification

The proposed granular space based model for pattern classification is illustrated in Fig. 1. The model has three steps of operation. A brief description and analysis of the advantages of each of the steps is made in the following.

The first step generates the class-dependent (CD) fuzzy granulated feature space of input pattern vector. For fuzzy granulation, **L** number of fuzzy sets are used to characterize the feature values of each pattern vector, where **L** is the total number of classes. We have considered π-type membership function (MF) for the

fuzzification purpose. Each feature is thus represented by **L** [0,1]-valued MFs representing **L** fuzzy sets or characterizing **L** fuzzy granules along the axis. The π-type MF explores the degree of belonging of a pattern into different classes based on individual features and the granules thus provide an improved class-wise representation of input patterns. The granules preserve the interrelated class information to build an informative granular space which is potentially useful for improved classification for the data sets with overlapping classes. The detail description and advantages of the method are given in Section 2.1.

In the granulation process, each feature value is represented with more than one membership values and thus the feature dimension increases. The increased dimension brings great difficulty in solving many tasks of pattern recognition, machine learning and data mining. This motivates for selecting a subset of relevant and non-redundant features. In this regard, we have used the neighborhood rough set [16,17] (NRS) based feature selection method in the second step of the proposed model (Fig. 1). The advantage in the use of NRS is that it can deal with both numerical and categorical data. NRS does not require any discretisation of numerical data and is suitable for the proposed fuzzy granulation of features. Further, the neighboring concept facilitates to gather the possible local information through neighbor granules that provide a better class discrimination information. Thus the combination of these two steps of operations can be a better framework for the classification of patterns in overlapping class environment. The proposed model thus takes the advantage of both class-dependent fuzzy granulation and NRS feature selection methods. Section 2.2.2 describes the detail procedure of NRS based feature selection. After the features are selected, we use a classifier as in the third step of Fig. 1 to classify the input pattern based on the selected features.

Along with the proposed model for pattern classification, we have described a new classifier performance measure based on the dispersion of classified patterns among classes. The dispersion measure is estimated from the confusion matrix that is obtained in the classification process. According to our definition, a smaller value of dispersion measure leads to a better classifier. The value decreases as the number of classes containing the misclassified patterns decreases. Thus, the dispersion measure of a class not only reflects its overlapping character with other classes, but also enables one to concentrate on the minimum number of classes, in order to get the misclassified patterns corrected with available higher-level information. This measure thus provides a better clue in the improvement of classifier's performance for individual classes. The detail description and significance of the measure is given in Section 3.1.

2.1. Class-dependent fuzzy granule generation

The class-dependent (CD) fuzzy granulated feature space is generated using fuzzy linguistic representation of pattern. Only the case of numeric features is mentioned here (features in descriptive and set forms can also be handled in this framework).

Let a pattern (object) **F** be represented by n numeric features and can be expressed as: $\mathbf{F}=[F_1,F_2,\ldots,F_n]$. Thus **F** is visualized as a point in n-dimensional vector space. Each feature is described in terms of its fuzzy membership values corresponding to **L** (=total number of classes) linguistic fuzzy sets. Thus, an n-dimensional pattern vector is expressed as (n x **L**)-dimensional vector and is given by

$$\begin{aligned}\mathbf{F}=[&\mu_1^1(F_1),\mu_2^1(F_1),\ldots,\mu_c^1(F_1),\ldots,\mu_{\mathbf{L}}^1(F_1);\\ &\mu_1^2(F_2),\mu_2^2(F_2),\ldots,\mu_c^2(F_2),\ldots,\mu_{\mathbf{L}}^2(F_2);\\ &\mu_1^n(F_n),\mu_2^n(F_n),\ldots,\mu_c^n(F_n),\ldots,\mu_{\mathbf{L}}^n(F_n)]\quad (c=1,2,\ldots,\mathbf{L}),\end{aligned}$$

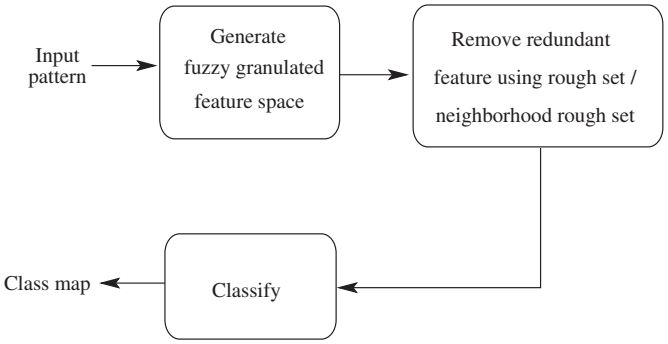

Fig. 1. Schematic flow diagram of the proposed model for pattern classification.

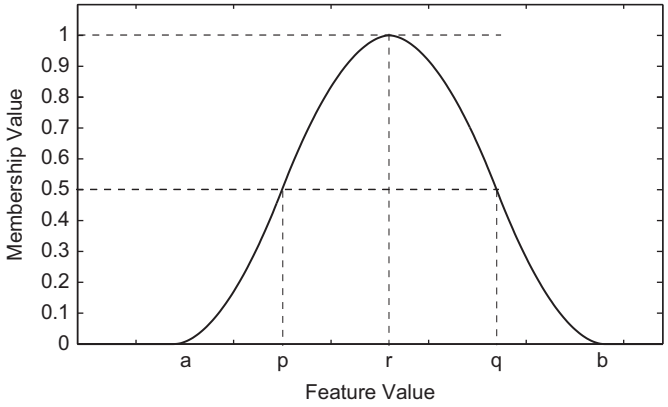

Fig. 2. π-Type membership function.

where $\mu_1^n(F_n),\mu_2^n(F_n),\ldots,\mu_c^n(F_n),\ldots,\mu_{\mathbf{L}}^n(F_n)$ signify the membership values of F_n to **L** number of fuzzy sets along the nth feature axis and $\mu(F_n)\in[0,1]$. It implies that each feature F_n is expressed separately by **L** number of MFs expressing **L** fuzzy sets. In other words, each feature F_n characterizes **L** number of fuzzy granules along each axis and thus comprising $\mathbf{L}^n$ fuzzy granules in an n-dimensional feature space. Zadeh's π-type MF [26,27] is used to characterize the fuzzy granules. The membership value $\mu_c^n(F_n)$ thus generated expresses the degree of belonging of nth feature to cth class of the pattern **F**. The π-type MF is given by

$$\begin{aligned}\mu(F_n;a,r,b)&=0,\quad F_n\le a\\ &=2^{N-1}[(F_n-a)/(r-a)]^N,\quad a<F_n\le p\\ &=1-2^{N-1}[(r-F_n)/(r-a)]^N,\quad p<F_n\le r\\ &=1-2^{N-1}[F_n-r)/(b-r)]^N,\quad r<F_n\le q\\ &=2^{N-1}[(b-F_n)/(b-r)]^N,\quad q<F_n<b\\ &=0,\quad F_n\ge b,\end{aligned}\tag{1}$$

where N is the fuzzifier of the MF, as shown in Fig. 2. The MF can be estimated with center at r and $r=(p+q)/2$, where p and q are the two crossover points. The membership values at the crossover points are 0.5 and at the center r, its value is maximum (i.e., 1). Assignment of membership value is made in such a way that training data gets a value closer to 1, when it is nearer to the center of MF and a value closer to 0.5 when it is away from the center. For the determination of the MF, we have considered the position of the center at *mean* point. The *mean* point is estimated as $r=mean(n)$ (i.e., average value of the data set for feature n) as the center, then the two crossover points p and q are estimated as $p=mean(n)-[\max(n)-\min(n)]/2$, and $q=mean(n)+[\max(n)-\min(n)]/2$, where min and max are the minimum and maximum value of the data set for feature n. Thus the feature-wise class membership of input pattern can be automatically

determined from the training data. In the π-type MF we have used the two boundary points (min and max values) as the crossover points of class membership function because these are the most ambiguous points in fuzzy set theory [26] in terms of belongingness to a class, or possessing some imprecise property represented by the set. Therefore the MF will have sensitivity if the training set has an outlier.

The above fuzzification process of input features generates the CD fuzzy granulated feature space in n-dimension. All together the granulated feature space contains L^n granules in n-dimension with fuzzy boundaries among them. For a better visualization of the granules generated by the proposed model, we have converted fuzzy membership values to the patterns to binary ones, i.e., fuzzy MFs to binary functions using α-cut. This is demonstrated in Fig. 3. Here 0.5-cut is used to obtain $4^2=16$ crisp granules for four classes, as an example, in two-dimensional feature space. This is explained further visually in Fig. 4 which generates both CD and CI fuzzy granules in two-dimensional feature space for a four-class data set. As described earlier, one fuzzy set corresponding to each class along a feature is used for CD granulation, whereas each feature (irrespective of the number of classes) is described by three fuzzy sets bearing linguistic properties low, medium and high, in case of CI granulation. As a result, eight and six granules are generated here for a pattern $X(f_1,f_2)$ of four-class data set using CD and CI granulation process, respectively.

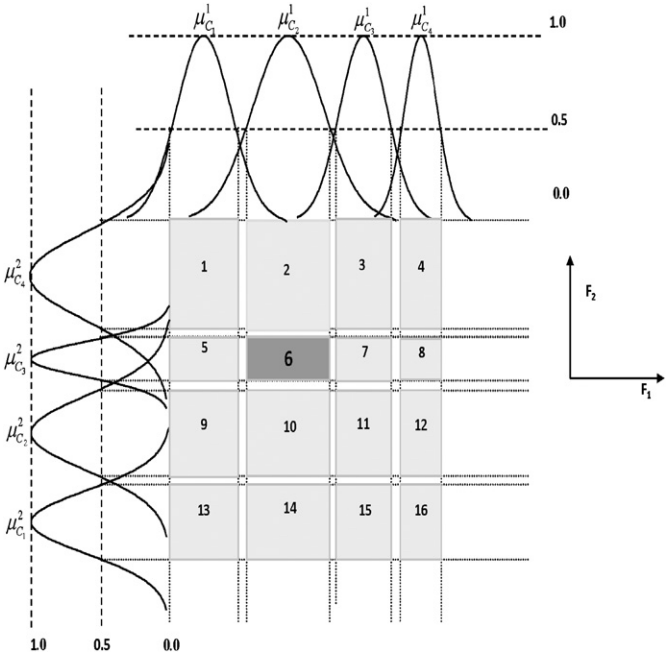

Fig. 3. Generation of crisp granules from class-wise (class-dependent) fuzzy (linguistic) representation of the features F_1 and F_2. The figure represents the granules for four overlapping classes. The shaded regions (16 nos.) indicate the granules. For example the region (granule no. 6) indicates a crisp granule obtained by α-cuts ($\alpha=0.5$ in present case) on the $\mu^1_{L_2}$ and $\mu^2_{L_3}$. The granules shape/size are variable in nature and depend on the overlapping nature of classes and class-wise feature distribution.

2.2. Feature selection

This section presents some preliminaries relevant to feature selection methods using rough sets (proposed by Pawlak) and neighborhood rough sets (NRS). The details of these theories may be referred to [5,16,17].

2.2.1. Rough sets (PaRs)

Pawlak's rough set (PaRS) theory [5] deals with vague concepts and creates approximate descriptions of objects for data analysis. It works with a pair of precise concepts, called *lower* and *upper* approximations. PaRS have been employed to remove redundant conditional features, while retaining their information content. The basic operation involved in PaRS is that it partitions the object space based on a feature set using some equivalence relation. The partition spaces thus generated are also known as *granules*, which become the elemental building blocks for data analysis.

A brief description of rough set (RS) theory used for feature selection is given here. Let an information system $IS=(U,A)$ be defined in terms of notions: U, the non-empty set of finite objects; A, the non-empty set of finite features, and $A=\{C\cup D\}$ where C and D are the set of conditional and decision feature, respectively. For any

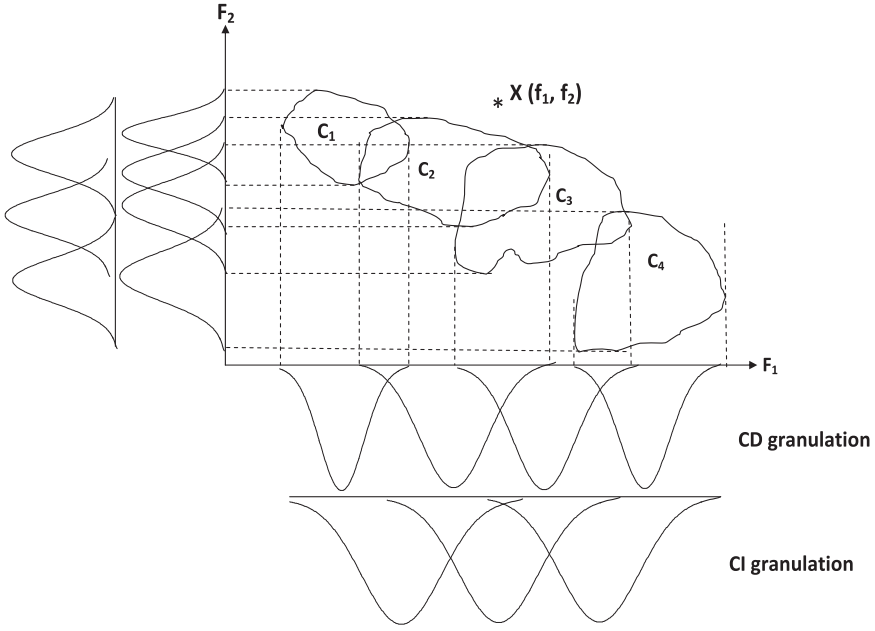

Fig. 4. Physical interpretation of fuzzy granule generation.

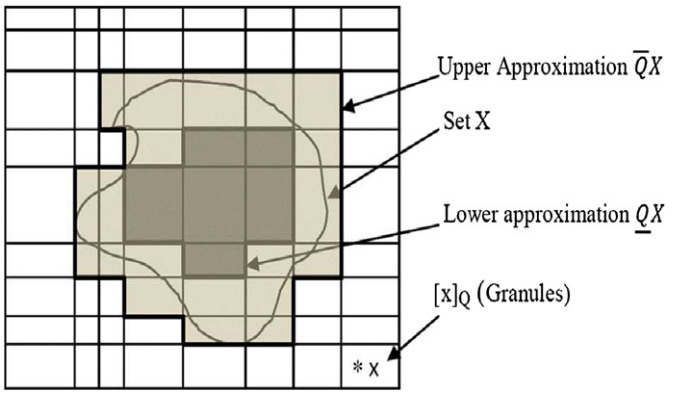

Fig. 5. Rough representation of a set with upper and lower approximations.

$Q \subseteq A$ and $X \subseteq U$, the set X can be approximated with the information available in Q by using lower ($\underline{Q}X$) and upper ($\overline{Q}X$) approximation of X, where $\underline{Q}X = \{x | [x]_Q \subseteq X\}$, $\overline{Q}X = \{x | [x]_Q \cap X \neq \phi\}$, and $[x]_Q$ denotes the equivalence class of the object $x \in U$ relative to I_Q (the equivalence relation). The objects in $\underline{Q}X$ can be certainly classified as members of X, while objects in $\overline{Q}X$ can only be classified as possible members of X on the basis of knowledge in Q. These are illustrated in Fig. 5 where the sets of dark-gray granules represent lower approximation, while those of both dark-gray and light-gray granules together denote upper approximation. The rough set thus defined with crisp set with rough representation. Let P and Q be two equivalence relations over U. A Q-positive region of P is the set of all objects from U which can be classified with certainty employing attributes from Q, and is defined as

$$POS_Q(P) = \bigcup_{X \in U/P} \underline{Q}X. \tag{2}$$

With the definition of positive region, degree of dependency of a set of features P on a set of features Q can be calculated as

$$k = \gamma_Q(P) = \frac{|POS_Q(P)|}{|U|}, \tag{3}$$

where $|\bullet|$ stands for the cardinality of the set.

The selection of features can be achieved through the comparison of equivalence relations generated by subsets of features. Features are removed such that the reduced set provides identical predictive capability of the decision feature(s) as that of original or unreduced set of features. With this concept, a measure of significance can be determined by evaluating the change in dependency when a feature is removed from the set. The higher the change in dependency, the more significant the feature is. Based on this significance a minimum element feature subset (reduct) is searched and located. Many attempts have been made for finding a reduct of an information system. The simplest solution for locating reducts is to generate all possible subsets and retrieve those with a maximum rough set dependency degree. However, this approach of finding solution is highly expensive for large data sets. For such cases, often one reduct instead of many is required to use for feature reduction. In this regard, the QUICKREDUCT algorithm described by Chouchoulas and Shen [28]is popularly used. The algorithm attempts to calculate a reduct without exhaustively generating all possible feature subsets. It starts with an empty set and adds one feature at a time that results in the increase of rough set dependency. The process goes on until it produces the maximum possible dependency value for a data set. The QUICKREDUCT algorithm is summarised with pseudocode, as shown in Algorithm 1.

Algorithm 1. QUICKREDUCT.

Input: C, the set conditional features; D, the set of decision feature
Output: R, the reduct, $R \subseteq C$
1: Initial the feature reduct $R = \phi$ and a temporary variable $T = \phi$
2: **do**
3: {
4: $T = R$
5: For every $a \in (C - R)$
6: {
7: **if** $\gamma_{R \cup \{a\}}(D) > \gamma_T(D)$, Then $T = R \cup \{a\}$
8: }
9: $R = T$
10: }
11: **until** $\gamma_R(D) = \gamma_C(D)$
12: **return** R

In the present study, we have used QUICKREDUCT algorithm for selecting features generated from the CD fuzzy granulation. The selected features are then used in a classifier for classifying the input pattern, as in the third step of Fig. 1.

2.2.2. Neighborhood rough sets (NRS)

As mentioned above the information system is denoted by $I = (U, A)$, where U (the universal set) is a non-empty and finite set of samples $\{x_1, x_2, \ldots, x_n\}$; $A = \{C \cup D\}$, where A is the finite set of features $\{a_1, a_2, \ldots, a_m\}$, C is the set of conditional features and D is the set of decision features. Given an arbitrary $x_i \in U$ and $B \subseteq C$, the neighborhood $\Phi_B(x_i)$ of x_i with given Φ, for the feature set B, is defined as [17]

$$\Phi_B(x_i) = \{x_i | x_j \in U, \Delta^B(x_i, x_j) \leq \Phi\}, \tag{4}$$

where Δ is a distance function.

$\Phi_B(x_i)$ in Eq. (4) is the neighborhood information granule centered with sample x_i. In the present study, we have used three **p**-norm distances in Euclidean space. These are Manhattan distance ($\mathbf{p}=1$), Euclidean distance ($\mathbf{p}=2$) and Chebychev distance ($\mathbf{p}=\infty$). The neighborhood granule generation is affected by two key factors such as the used distance function Δ and parameter Φ. The first one determines the shape and second controls the size of neighborhood granule. For example, with Euclidean distance the parameter Φ acts as the radius of the circle region developed by Δ function. Both these factors play important roles in neighborhood rough sets (NRS) and can be considered as to control the granularity of data analysis. The significance of features vary with the granularity levels. Accordingly, the NRS based algorithm selects different feature subsets with the change of Δ function and Φ value. In the present study, we have analyzed the effects of three **p**-norm distances for a variation of Φ values, and selected the best one based on the performance with the present data sets. However, optimal parameter values can be obtained through an optimization technique, e.g., genetic algorithm.

Thus each sample generates granules with a neighborhood relation. For a metric space $\langle U, \Delta \rangle$, the set of neighborhood granules $\{\Phi(x_i) | x_i \in U\}$ forms an elemental granule system, that covers the universal space rather than partitions it as in case of PaRS. A pictorial view of the process of granule generation (as an example) using both PaRS and NRS is shown in Fig. 6.

Let $X = \{a,b,c,d,e,f\}$ be the universal set of five elements (Fig. 6). Partitioning and covering of set X for generating granules are made as $X1 = \{\{a,b\}, \{c,d\}, \{e,f\}\}$ and $X2 = \{\{a,b\}, \{a,c,d\}, \{a,b,e,f\}\}$, respectively. A partition of the set X is a division of X into non-overlapping and non-empty "parts" or "blocks" or "cells" that accommodate all the elements of X. Equivalently, a set $X1$ of non-empty sets is

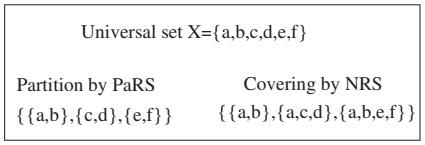

Fig. 6. Example of granule generation using PaRS and NRS.

a partitions of X, if the intersection of any two distinct elements of $X1$ is empty. On the other hand, a covering of a set X results into overlapping and non-empty "parts" that accommodate all the elements of X. That means a set $X2$ of non-empty sets is a covering of X, if the intersection of any two distinct elements of $X2$ is not necessarily empty. It is noted that the partition of space generated by PaRS can be obtained from NRS with covering principle, while the other way round is not possible. Moreover, a neighborhood granule degrades to an equivalent class for $\Phi=0$. In this case, the samples in the same neighborhood granule are equivalent to each other and the neighborhood rough set model degenerates to Pawlak's rough set. Thus NRS can be treated as a generalized case of PaRS.

A neighborhood information system can de denoted by $NIS=(U,A,N)$ when a feature in the system generates a neighborhood relation on the universal set U. The set of features $A=\{C\cup D\}$, where C and D are the set of conditional and decision features, respectively. Let $X_1,X_2,\ldots,X_N$ be the object subsets with decisions 1 to N and $\delta_B(x_i)$ be the neighborhood information granule generated by feature set $B\subseteq C$.. The lower and upper approximation of decision D with respect to features B are defined as

$$\underline{N_B}D=\bigcup_{i=1}^{N}\underline{N_B}X_i,\quad \overline{N_B}D=\bigcup_{i=1}^{N}\overline{N_B}X_i, \tag{5}$$

where

$$\underline{N_B}X=\{x_i\,|\,\delta_B(x_i)\subseteq X,x_i\in U\},\quad \overline{N_B}X=\{x_i\,|\,\delta_B(x_i)\cap X\neq\emptyset,x_i\in U\}.$$

The decision boundary region of D with respect to features B is defined as

$$BN(D)=\overline{N_B}D-\underline{N_B}D.$$

The lower approximation of the decision is defined as the union of the lower approximation of each decision class. The lower approximation of the decision is called the positive region of the decision, denoted by $POS_B(D)$. $POS_B(D)$ is the subset of objects whose neighborhood granules consistently belong to one of the decision classes. A sample in the decision system belongs to either the positive region or the boundary region of decision. Therefore, the neighborhood model divides the samples into two subsets: positive region and boundary region. Positive region is the set of samples which can be classified into one of the decision classes without uncertainty, while boundary region is the set of samples which cannot be determinately classified.

The dependency degree of decision feature D on condition feature set B in a neighborhood information system $\langle U,C\cup D,N\rangle$ with distance function Δ and neighborhood size Φ is defined as

$$\gamma_B(D)=\frac{|POS_B(D)|}{|U|}, \tag{6}$$

where $|\bullet|$ denotes the cardinality of a set. $\gamma_B(D)$ is the approximation ability of B to D. For $POS_B(D)\subseteq U$, we have $0\le\gamma_B(D)\le 1$ and D depends completely on B, and the decision system is consistent in terms of Δ and Φ. For $\gamma_B(D)=1$, D depends on B in the degree of γ. The dependency function measures the approximation power of a condition feature set. Hence it can be used to determine the significance of a subset of features (normally called reduct). Significance (SIG) of a subset of features is calculated with the change of dependency. Two types of features selection such as backward search and forward search can be made and accordingly the SIG factors are measured. The backward search starts with the reduct containing the whole conditional features and in each step a feature is removed from it. On the other hand, forward search begins with an empty reduct and in each step a feature from conditional features set is added to it.

Algorithm 2. Forward greedy search.

Input: C, the set conditional features; D, the set of decision feature
Output R, the reduct, $R\subseteq C$
1: $R\leftarrow\{\}$
2: **For** every $a_i\in(C-R)$
3: Compute $\gamma_{R\cup a_i}(D)=\frac{|POS_{B\cup a_i}(D)|}{|U|}$
4: Compute significance $SIG(a_i,R,D)=\gamma_{R\cup a_i}(D)-\gamma_R(D)$
5: **end**
6: Select the feature a_k satisfying $SIG(a_k,R,D)=\max_i(SIG(a_i,R,D))$
7: **if** $SIG(a_k,R,D)>\epsilon$, where ϵ is the stopping parameter
8: $R\leftarrow R\cup a_k$
9: go to step '2'
10: **else**
11: return R
12: **end if**

With these searching processes, many sets of reducts can be obtained based on the significance factor and any of them will work for the feature reduction task. In this regard Hu et al. [17] described a forward greedy search (FGS) algorithm for feature selection using NRS. Algorithm 2 summarises the pseudocode of FGS algorithm. Algorithm 2 starts with an empty set R of attribute. One feature is added to the set R which makes the increment of dependency (SIG value) maximal in each step. The algorithm stops if SIG falls below a small value ϵ by adding any new feature into the attribute subset R [17]. In the present study, we have used the forward greedy search algorithm for the selection of features in the proposed rough-fuzzy granulation based model for classification. The reason for choosing the forward search method is that it requires less computational effort than the backward search method. This is because, to obtain the minimal feature subsets (reduct), the forward method starts with an empty set of features whereas backward method starts with the whole set of features. The selected features are then fed to a classifier for classifying the input pattern, as in the third step of Fig. 1.

3. Performance measurement indexes

3.1. Proposed dispersion measure

The performance of a classifier is analyzed here with respect to its confusion matrix (CM). The nature of distribution of the classified patterns among different classes reflects the overlapping characteristics of various regions. With this notion, the class-wise dispersion measure of a classifier can be defined that quantifies the nature of dispersion of the misclassified patterns into different classes.

Let Table 1 represent the CM of a classifier for a data set having three classes (L_1, L_2 and L_3) with 10 samples in each class. The rows of CM correspond to the true/reference, and columns as assigned/estimated class labels. Elements in each row thus represent the distribution of the classified patterns for a particular class

of interest. For example, $E_{12}=2$ indicates that two samples out of 10 from class 1 are assigned to class 2.

The dispersion score (DS) for ith class may be defined as

$$(DS)_i=\left[1-\left(\lambda_1\frac{V_i}{V^i_{max}}+\lambda_2\frac{Z_i}{L-1}\right)\right],\quad i,=1,2,\ldots,L \text{ (number of classes)}, \tag{7}$$

where V_i is the variance of the elements in ith row of CM, e.g., $V_{i,i=1}$ of CM (Fig. 1)$=variance[E_{11}, E_{12}, E_{13}]$, V^i_{max} is the maximum variance corresponding to ith row, which is obtained when all the patterns are correctly classified to the ith class, Z_i is the number of element(s) in ith row of CM with zero value, $Z_i\le(L-1)$, e.g., $Z_{i,i=1}$ of CM (Fig. 1)$=1$, λ_1,λ_2 are the weight factors, and $\lambda_1+\lambda_1=1$.

The first part of Eq. (7) is the normalized variance for the class of interest and quantifies the nature of distribution of the classified patterns among the classes. The second part of Eq. (7) is the normalized count for the number of classes where no patterns are classified. We have combined both the aspects to reflect the overlapping characteristic of one class with others. Hence Eq. (7) provides a measure quantifying the class-wise distribution of the classified patterns for a classifier. The measure can be viewed as an index in evaluating the class-wise performance of a classifier. In Eq. (7), the weight factors λ_1 and λ_2 can be assigned according to the requirement in hand. That means, more is the importance to an aspect, higher is the weight value.

According to the dispersion measure, less is the DS value better is the agreement. DS$=0$ indicates perfect agreement between the two observers (true and estimated), i.e., all the test patterns of that class are correctly classified. If the patterns are misclassified to (i.e., confused into) minimum number of classes, DS value would be less. Therefore with a given number of overlapping classes for a particular class of interest in the feature space, lower DS value is desirable. Lower value would also facilitate one to focus on less number of confused classes in order to get some of the misclassified patterns corrected with available higher level information. The characteristics and significance of the measure can be further illustrated with examples in the following section.

Table 1
A 3×3 confusion matrix, where E_{ij} denotes the number of patterns actually from ith class, classified to jth class.

	Assigned classes		
	L_1	L_2	L_3
True classes			
L_1	$E_{11}=8$	$E_{12}=2$	$E_{13}=0$
L_2	$E_{21}=1$	$E_{22}=7$	$E_{23}=2$
L_3	$E_{31}=1$	$E_{32}=0$	$E_{33}=9$

3.1.1. Characteristics and significance of dispersion measure

Let one of the six classes of a synthetic remote sensing image data (Section 4.1.7) consists of 5000 number of patterns. The dispersions of these patterns by 10 different classifiers based on the aspects 1, 2 and both (Eq. (7)) are represented by 10 rows as shown in Fig. 7. According to the first aspect (defined in the first part of Eq. (7)), classifier 1 is ranked as the best and classifier 9 as the worst in performance scale. This is because all the patterns are classified to one class (class 1, the correct class) with classifier 1, whereas more number of misclassified patterns are distributed among the classes with classifier 9. In a comparison between classifiers 4 and 6, classifier 6 is ranked higher in spite of 4505 patterns being classified to one class (class 1) whereas it is 4516 with the classifier 4. This is because 4947 patterns are classified in two classes (classes 1 and 2) with classifier 6, whereas it is 4750 with classifier 4. Similar comparison can be made with classifiers 3 and 9, where the earlier classifier is superior to latter.

With a comparison of classifiers' performance based on aspect 2, classifier 1 is superior because here five classes have zero classified patterns, which is the highest number of possible empty classes for this example. Classifiers 8, 9 and 10 are seen to be the worst performer because none of the classes for them is empty. In another comparison, classifier 3 is superior to classifier 4 because two classes are empty with the former compared to one class with the latter. However, this ranking is changed when both the aspects are required to be satisfied together.

While considering both the aspects together, we have given more importance to aspect 1 compared to aspect 2, i.e., $\lambda_1=0.75$ and $\lambda_2=0.25$. The reason is that the variance estimation of pattern distribution among the classes mainly takes care of the first aspect and also to some extent for the second one. That means aspect 1 by default searches for the classes with zero number of classified patterns. If we combine both the aspects then the combined effect would lead to finding a more justifiable dispersion measure of classified patterns. The following examples taken from Fig. 7 will justify the above discussion. It is observed from Fig. 7 that with aspect 1, classifier 8 is ranked second in descending order of performance scale and is superior to classifier 2 which is ranked fourth. However with the combined aspect, classifier 2 secured second position and classifier 8 third position because the dispersion score of 0.1500 with classifier 2 compared to 0.2500 with classifier 8, obtained with aspect 2, is added to the combined

Classifier	Classes						Dispersion score for different aspects with $\lambda_1=0.75$ and $\lambda_2=0.25$					
	1	2	3	4	5	6						
							Aspect 1 = $0.75-\lambda_1\frac{V_i}{V^i_{max}}$		Aspect 2 = $0.25-\lambda_2\frac{Z_i}{C-1}$		Both	
							Score	Rank (descending)	Score	Rank (descending)	Score	Rank (descending)
1	5000	0	0	0	0	0	0.0000	1	0.0000	1	0.0000	1
2	4507	425	47	21	0	0	0.1621	4	0.1500	2	0.3121	2
3	4005	674	218	103	0	0	0.3041	9	0.1500	2	0.4541	9
4	4516	234	99	85	66	0	0.1631	6	0.2000	3	0.3631	5
5	4411	397	96	42	54	0	0.1934	7	0.2000	3	0.3934	7
6	4505	442	33	8	12	0	0.1623	5	0.2000	3	0.3623	4
7	4310	387	210	45	48	0	0.2241	8	0.2000	3	0.4241	8
8	4708	86	54	57	48	47	0.1014	2	0.2500	4	0.3514	3
9	4020	487	210	146	95	42	0.3069	10	0.2500	4	0.5569	10
10	4606	188	48	57	53	48	0.1346	3	0.2500	4	0.3846	6

Fig. 7. Typical example for the dispersion measure with10 classifiers.

score. That means according to the combined aspect, classifier 2 is superior to classifier 8, which is intuitively acceptable. In another comparison between classifiers 7 and 8, the latter is superior to former both with aspect 1 and in combined case, although the former is superior to latter with aspect 2. In spite of all the patterns being confined to five classes with classifier 7 and six classes with classifier 8, the concept of the number of classes with zero number of classified patterns did not support classifier 7 for superiority unlike in the previous example. This is reasonable because classifier 8 contained 4708 patterns in one class whereas classifier 7 needed more than two classes (i.e., classes 1 and 2) to accommodate equivalent number of patterns. Similar comparison can be made with classifiers 3 and 4. Hence the first example justified the reason of combining the two aspects whereas the second and third examples justified for giving more importance to aspect 1 compared to aspect 2. Based on these findings we see that dispersion score (DS) measure of a class reflects its overlapping character with other classes, and it can be used as an index for classifier's efficiency.

Apart from providing an index of class-wise performance, the DS measure provides a helpful clue for improving the class-wise performance with additional (higher-level) information. Smaller the DS value (i.e., when the misclassified patterns are mostly confined into minimum number of classes), larger the possibility that some of the misclassified patterns from the neighboring class(es) may get rectified with the higher-level (e.g., syntactic, semantic, contextual, etc.) information. Let us consider the problem of analyzing remote sensing images where the land cover classes are highly overlapping. Before determining the classification accuracy of a particular class, it may be helpful and appropriate to know the other classes overlapped with it. For example, consider the scene of a bridge over a river. There is a possibility that some of the bridge pixels would be misclassified as water body. In that case if the misclassified pixels are only confined into water body as indicated by lower DS value, then with the help of higher-level information like, a bridge must be connected to roads on either side, some of the said pixels misclassified as water body can be easily corrected as bridge pixels; thereby enabling the detection of the bridge structure.

3.2. Overall classification accuracy, Precision, Recall and kappa coefficient

Along with the aforesaid DS measure, we have used other indexes such as classification accuracy, Precision, Recall and kappa coefficient for completely labeled data sets, and β and Davies–Bouldin (DB) for partially labeled data sets, while measuring the classifier's performance. Brief descriptions of these indexes are provided in this following.

To examine the practical applicability of proposed model for completely labeled data sets various performance measures are used. These are percentage of overall classification accuracy (PA), Precision, Recall and kappa coefficient (KC). The PA value is the percentage of samples that are correctly classified and can be evaluated from confusion matrix (CM). In the present study, we have considered the significance of CM with respect to individual class. Sometimes a distinction is made between errors of omission and errors of commission, particularly when only a small number of class type is of interest. Thus interpreting a CM from a particular class point of view, it is important to notice that different indications of class accuracies will result differently according to whether the number of correct patterns for a class is divided by the total number of true (reference) patterns for the class or the total number of patterns the classifier features to the class. The former is normally known as Recall and the latter as Precision [22]. The following example illustrates the calculation of these measures. Let a data set be represented with three classes. The corresponding CM generated by a classifier is depicted in Table 2.

Percentage of overall accuracy (PA) is calculated from the CM (Table 2) as

$$\mathrm{PA} = \frac{A+E+I}{T}.$$

Similarly, for class L_1, Precision and Recall are calculated as

$$\mathrm{Precision} = \frac{A}{X1}, \quad \mathrm{Recall} = \frac{A}{Y1}.$$

Note that the overall classification accuracy does not provide the class-wise agreement between the true and estimated class labels and the Precision and Recall measures give the results for individual class only. To get an overall class agreement based on the individual class accuracy, we have used kappa coefficient (KC) [23] estimation to validate the superiority of the classifiers effectively. The KC measure was introduced by the psychologist Cohen [23] and adapted for accuracy assessment in the remote sensing field by Congalton and Mead [29]. The KC and classification accuracy are not proportional, that means a good percentage of accuracy may lead to a poor KC value, because it provides the measurement of class-wise agreement between the true and estimated class labels. Higher is the coefficient value better is the agreement of the estimated data with the true one. The KC value is estimated from a CM as follows:

$$\mathrm{KC} = \frac{M\sum_{i=1}^{r} Y_{ii} - \sum_{i=1}^{r}(Y_{i+} \times Y_{+i})}{M^2 - \sum_{i=1}^{r}(Y_{i+} \times Y_{+i})}, \tag{8}$$

where r is the number of rows in the error matrix, Y_{ii} is the number of observations in row i and column i, Y_{i+} is the total observation in row i, Y_{+i} is the total observation in column i, and M is the total number of observations included in the matrix.

A KC value (>0) indicates the amount of agreement between the two observers (true and estimated). A value of 1 indicates perfect agreement (when all the values are falling on the diagonal) [23].

Table 2
Typical example for the calculation of different measures.

	True classes			Total
	L_1	L_2	L_3	
Assigned classes				
L_1	A	B	C	$A+B+C=X1$
L_2	D	E	F	$D+E+F=X2$
L_3	G	H	I	$G+H+I=X3$
Number of true patterns	$A+D+G=Y1$	$B+E+H=Y2$	$C+F+I=Y3$	$X1+X2+X3=Y1+Y2+Y3=T$

3.3. β index

The β index has been defined by Pal et al. in [24], for assessment of image segmentation quality. β is defined as the ratio of the total variation and within-class variation as

$$\beta = \frac{\sum_{i=1}^{L}\sum_{j=1}^{M_i}(\mathbf{x}_{ij}-\overline{\mathbf{x}})^2}{\sum_{i=1}^{L}\sum_{j=1}^{M_i}(\mathbf{x}_{ij}-\overline{\mathbf{x}}_i)^2}, \tag{9}$$

where $\overline{\mathbf{x}}$ is the mean grey value of the image pixels (pattern vector), M_i is the number of pixels in the ith ($i=1,2,\ldots,L$) class, $\mathbf{x}_{ij}$ is the grey value of the jth pixel ($j=1,2,\ldots,M_i$) in class i, and $\overline{\mathbf{x}}_i$ is the mean of M_i grey values of the ith class. Since the numerator is constant for a given image, β value is dependent only on the denominator. The denominator decreases with increase in homogeneity within the class for a fixed number of classes ($\mathbf{L}$). Thus for a given image and given number of classes, the higher the homogeneity within the classes and lower the homogeneity between classes, the higher would be the β value.

3.4. Davies–Bouldin index (DB)

Davies–Bouldin (DB) index for cluster validation has been defined in [25]. However, here we are using the index for validating our classification results on partially labeled data sets. The idea behind DB index is that for a good partition inter-cluster separation as well as intra-cluster homogeneity and compactness should be high. The DB index is based on the evaluation of some measure of dispersion S_i within the ith cluster and the distance between the prototypes of clusters i and j. The DB index is defined as

$$\mathrm{DB} = \frac{1}{K}\sum_{i=1}^{K} R_{i,qt}, \tag{10}$$

where K is the number of clusters/classes and $R_{i,qt} = \max_{j,j\neq i}[(S_{i,q}+S_{j,q})/d_{ij,t}]$. $S_{i,q}$ is the qth root of qth moment of the points in cluster i with respect to their mean or centroid. $d_{ij,t}$ is the Minkowski distance of order t between the centroids that characterize the extracted classes i and j. The smaller the DB value the better is the partitioning.

4. Description of the data sets used

In the present study, we have chosen seven completely labeled data sets (listed in Table 3) including synthetic remote sensing image and two partially labeled real remote sensing images collected from satellite. A short description for these data sets is provided below.

4.1. Completely labeled data sets

4.1.1. VOWEL data

VOWEL data [30] is a set of Indian Telegu vowel sounds in consonant–vowel–consonant context uttered by three male speakers in the age group of 30–35 years. As a whole the data set consists of 871 patterns. It has three features and six classes /∂/, /a/, /i/, /u/, /e/ and /o/ with 72, 89, 172, 151, 207 and 180 samples, respectively. The data set (depicted in Fig. 11 with two dimensions for ease of understanding) has three features: F_1, F_2, F_3 corresponding to the first, second, and third vowel formant frequencies obtained through spectrum analysis. The classes are highly overlapping and possess ill-defined boundaries.

Table 3
Brief description of the data sets used in the present study.

Sl. no.	Name of the data set	Number of classes	Number of features	Number of patterns
1	VOWEL	6	3	871
2	SATIMAGE	6	4	6435
3	WAVEFORM	3	21	5000
4	CALDONAZZO	6	7	3884
5	PHONEME	2	5	5404
6	PAGE-BLOCK	5	10	5473

4.1.2. SATIMAGE data

The SATIMAGE data [31] is generated from Landsat Multi-Spectral Scanner image data. The data patterns used for the present investigation is a sub-area of a scene of 82×100 pixels. Each pixel value contains information from four spectral bands. The aim is to predict six different land cover classes present in the data set. The data set contains 6435 patterns with 36 features (4 spectral bands $\times$ 9 pixels in neighborhood). In our experiment we have used four features only as recommended by the database designer [31].

4.1.3. WAVEFORM data

WAVEFORM data [32] consists of three classes of waves with 21 number of features. Each class of the data set is generated from a combination of base waves. All the features are corrupted with noise (mean 0, variance 1). The data set contains 5000 patterns. Class distribution of the patterns present in the data set is made with 33% for each class.

4.1.4. CALDONAZZO data

CALDONAZZO data is obtained from multispectral scanner satellite image. The data patterns used in the present study is a sub-area of a scene of 881 $\times$ 928 pixels. Each pixel value contains information from seven spectral bands. The data set contains 3884 patterns with the information of six different land cover classes.

4.1.5. PHONEME data

The aim of this data is to distinguish between nasal and oral vowels (two classes) [31]. It contains vowels coming from 1809 isolated syllables (for example, pa, ta, pan, etc.). The amplitudes of five first harmonics are chosen as features to characterize each vowel. The data set has 5404 patterns.

4.1.6. PAGE-BLOCK data

The problem involved in this data [32] is to classify the blocks of a page layout of a document that has been detected by a segmentation process. This is an essential step in document analysis in order to separate text from graphic areas. The five classes are text (1), horizontal line (2), picture (3), vertical line (4) and graphic (5). PAGE-BLOCK data set has 10 features and five classes with 5473 patterns.

4.1.7. Synthetic image

A four-band synthetic image (size 512×512) has been generated with six major land cover classes similar to the IRS-1A image. Fig. 8(a) shows the synthesized image in the near infrared range (band-4). Different classification models are tested on various corrupted versions of the synthetic images. The synthetic image is corrupted with Gaussian noise (zero mean and standard deviation (σ) $=1,2,\ldots,6$) in all four bands. Fig. 8(b) shows the noisy version of the original image with $\sigma=2$, as an example.

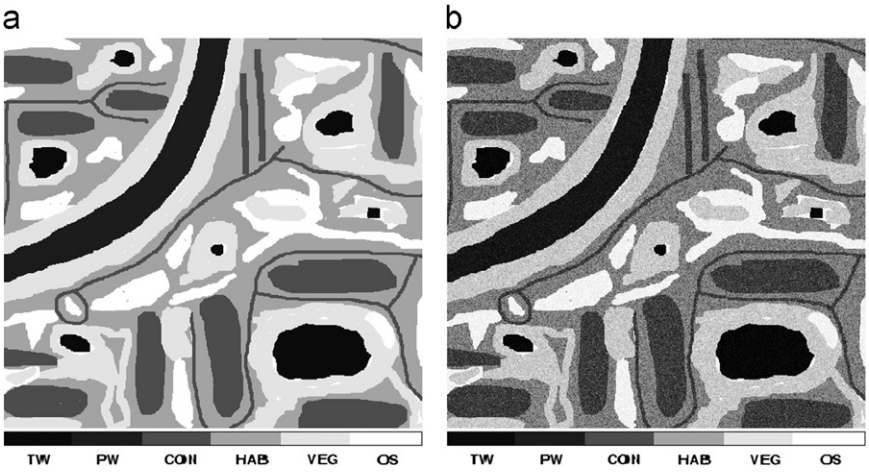

Fig. 8. Synthetic image (band-4). (a) Original and (b) noisy ($\sigma=2$).

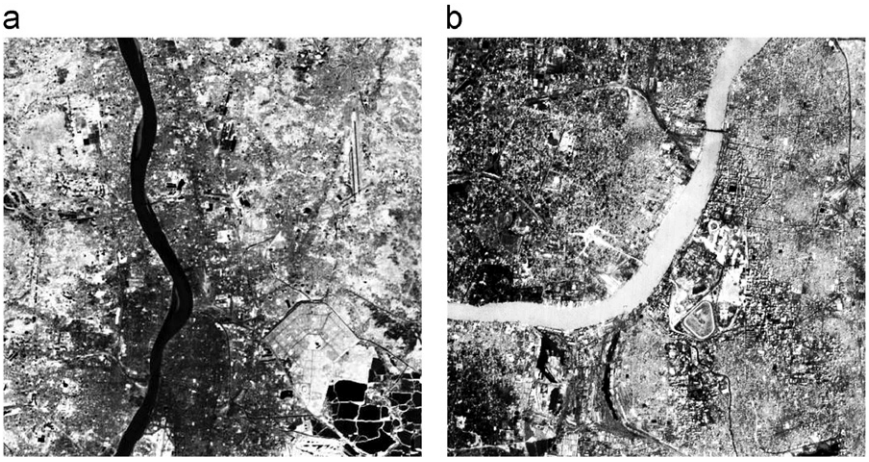

Fig. 9. (a) IRS-1A (band-4) enhanced image and (b) SPOT (band-3) enhanced image.

4.2. Partially labeled data sets

4.2.1. IRS-1A image

The IRS-1A image (size 512×512) is obtained from Indian Remote Sensing Satellite [33]. We have used the image taken from the Linear Imaging Self Scanner with spatial resolution of 36.25 m $\times$ 36.25 m and wavelength range of $0.45-0.86$ μm. The whole spectrum range is decomposed into four spectral bands namely, blue, green, red and near infrared corresponding to band-1, band-2, band-3 and band-4, respectively. Since the image is of poor illumination, we have presented the enhanced image (band-4) in Fig. 9(a) for the convenience of visualizing the content of the image. However, the algorithms are implemented on actual (original) image. The image in Fig. 9(a) covers an area around the city of Calcutta, India, in the near infrared band having six major land cover classes: *pure water* (PW), *turbid water* (TW), *concrete area* (CA), *habitation* (HAB), *vegetation* (VEG) and *open spaces* (OS).

4.2.2. SPOT image

The SPOT image (size 512×512) shown in Fig. 9(b) (enhanced image (band-3)) is obtained from SPOT satellite (Systeme Pour d'Observation de la Terre). The image used here has been acquired in the wavelength range of $0.50-0.89$ μm. The whole spectrum range is decomposed into three spectral bands namely, green (band-1), red (band-2) and near infrared (band-3) of wavelengths $0.50-0.59$ μm, $0.61-0.68$ μm, and $0.79-0.89$ μm, respectively. This image has a higher spatial resolution of 20 m $\times$ 20 m. We have considered the same six classes as in the case of IRS-1A image. Also, similar to IRS-1A image, the classification models are applied on the actual (not enhanced) image.

5. Results and discussion

Fuzzy information granulation has come up with an important concept of fuzzy set theory. Many attempts have been made in the construction of fuzzy granule and used for various image processing and pattern recognition applications [14,34–36]. In a similar attempt, recently, Fu et al. [37] has described a method of fuzzifing the artificial filters with some fuzzy decision rules and aggregation operators. However, the selection of aggregation operators or rules, as made above, is difficult in any decision making process and highly data dependent, which finally makes the approach empirical. Further, the task becomes cumbersome for the data sets with highly overlapping classes.

Our approach defers in the fuzzification process of the input, where the information extraction is data independent. The decision process collects the class dependent information from the available features of the data sets, and finally selects the likely optimum features using rough set theory for the classification task. In the present investigation we have compared the performance of the

proposed model with different combinations of fuzzy granulation and rough feature selection methods. The class-dependent (CD) granulation method is also compared with class-independent (CI) based method. For CI based granulation, the whole feature space is used for granule generation irrespective of classes. Each feature of the pattern is represented by three fuzzy sets for characterizing three fuzzy granules along each axis; thereby providing 3^n fuzzy granules in an n-dimensional feature space.

Five different combinations of classification models using granular feature space and feature selection methods those are considered for performance comparison as mentioned below. Patterns with its original feature representation are fed as input to these models.

- Model 1: k-nearest neighbor (k-NN with $k=1$) classifier.
- Model 2: CI fuzzy granulation + Pawlak's rough set (PaRS) based feature selection + k-NN (with $k=1$) classifier.
- Model 3: CI fuzzy granulation + neighborhood rough set (NRS) based feature selection + k-NN (with $k=1$) classifier.
- Model 4: CD fuzzy granulation + PaRS based feature selection + k-NN (with $k=1$) classifier.
- Model 5: CD fuzzy granulation + NRS based feature selection + k-NN (with $k=1$) classifier.

Apart from the performance comparison with different quantitative measures for both completely and partially labeled data sets, the efficacy of the proposed model of rough-fuzzy granulation and feature selection is also justified with the following types of analyses. However, the experimental results with these analyses are provided only for VOWEL data because similar trend of comparative performance is observed for the remaining data sets. The performance measure of models in terms of receiver operating characteristic (ROC) plot for the binary classification task of PHONEME data set is also made. ROC is a graphical plot of the false positive rate (1, specificity or 1, true negative rate) vs. sensitivity, or true positive rate, for a binary classifier system.

- Variation of classification accuracy with different values of parameter Φ and distances used in NRS based feature selection for optimal value selection.
- Performance comparison in terms of total computation time (T_c).
- Precision and Recall based analysis.
- Performance evaluation in terms of dispersion measure of different classes.
- Performance comparison of the proposed model with other classifiers such as k-NN with $k=3$ and 5, maximum likelihood (ML) classifier and multi-layered perceptron (MLP).
- Performance comparison of models in terms of β and Davies–Bouldin (DB) indexes.
- Statistical significance test called CHI-SQUARE for the proposed model.
- Performance comparison of the models with principal component analysis based feature reduction models.

5.1. Classification of completely labeled data sets

Selection of the training and test samples for all classes in case of completely labeled data sets including synthetic image has been made after splitting the whole data set into two parts as training and test sets. We have taken 10%, 20% and 50% as training set and the rest 90%, 80% and 50% are considered as test data. Training set are selected randomly and an equal percent of data is collected from each class. We repeat each of these splitting sets for 10 times and the final results are then averaged over them.

5.1.1. VOWEL data

The classification results for this data set with five different models using k-NN classifier ($k=1$) are depicted in Table 4 for three different percentages of training sets. In the present experiment, we have compared the classification performances with respect to three different aspects. These are performance based on (i) granulated and non-granulated feature space, (ii) class-dependent (CD) and class-independent (CI) fuzzy granulation, and (iii) Pawlak's rough sets (PaRS) and neighborhood rough sets (NRS) based feature selection. As described in Section 2.2.2, performance comparison with the NRS method of feature selection depends on the distance function Δ and parameter Φ of the neighborhood granules.

In the present study we have analyzed the performance of model 5 in terms of the variation of Δ and Φ for 20% training set of VOWEL data. We plotted the classification accuracy and the number of selected features (Fig. 10) for three $\mathbf{p}$-norm distances for a variation of Φ values ([0,1]) in Euclidean space. These are Manhattan distance ($\mathbf{p}=1$), Euclidean distance ($\mathbf{p}=2$) and Chebychev distance ($\mathbf{p}=\infty$).

As described earlier, variation of Φ values results in the change of granularity levels, and significance of attributes. Accordingly, the NRS based algorithm selects different feature subsets for different Φ values. The subsets may or may not contain the same features and lead to a possible variation in the number of features. Also there is a possibility of selecting very few/none of the features in the search process if the increment in dependency introduced with every new feature is not satisfactory. This situation normally comes when a higher value of Φ results in the construction of a large granule that accommodates more neighbors, thereby increasing the possibility of lowering the ratio of number of relevant features to irrelevant features. This is evident in Fig. 10(b) demonstrating the variation of the number of selected rough set features with Φ for all types of distances.

It is observed from Fig. 10(a) that the classification accuracy varies with Φ in a similar manner for all types of distances. With the increase of Φ value the accuracy increases at first, attains a fairly constant value for a moderately wide range of Φ, and then decreases. This is also justified from Fig. 10(b). For all the distances, the highest accuracy is obtained roughly for $\Phi=[0.2,0.5]$ with maximum value for Euclidean distance. Beyond $\Phi=0.5$, the neighborhood rough set based classification model, as discussed above, does not get enough relevant features to distinguish patterns and the classification

Table 4
Performance comparison of models using k-NN classifier ($k=1$) with VOWEL data ($\mathbf{p}=2$, $\Phi=0.45$).

Model	10% of training set			20% of training set			50% of training set		
	PA	KC	T_c (s)	PA	KC	T_c (s)	PA	KC	T_c (s)
1	73.240	0.7165	0.146	75.150	0.7212	0.124	77.560	0.7256	0.101
2	76.010	0.7502	0.236	77.010	0.7704	0.223	79.030	0.7721	0.214
3	77.870	0.7621	0.263	78.810	0.7853	0.252	80.790	0.7901	0.244
4	81.030	0.8008	0.365	81.370	0.8165	0.351	82.110	0.8202	0.345
5	83.750	0.8102	0.381	83.960	0.8253	0.378	84.770	0.8301	0.354

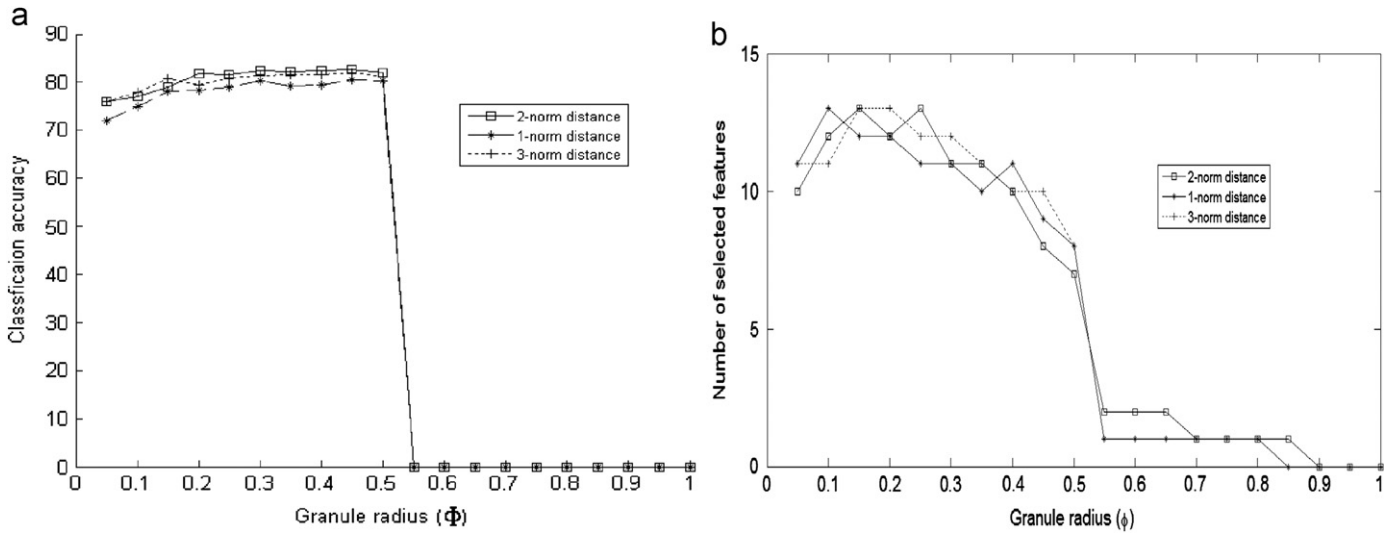

Fig. 10. Variation of (a) classification accuracy and (b) number of features for model 5 with size parameter Φ for three **p**-norm distances (20% training set).

accuracy falls. It is interesting to note that, although the numbers of selected features are different when Φ takes values in the interval [0.2,0.5], they are producing similar classification performance. This indicates that the Φ value should be varied in [0.2, 0.5] to find the minimal subset of features with similar (highest) classification accuracy. Beyond 0.5, finding that subset is drastically affected. Accordingly, for presenting the results for the remaining data sets, we have taken $\mathbf{p}=2$ (Euclidean distance) and $\Phi=0.45$.

In a comparative analysis, it is observed from Table 4 that for all percentages of training sets the classifiers' performance measured with percentage of accuracy (PA) values is more encouraging for models using granulated feature space. For example, with 10% of training set model 1 provided PA value of 73.24, whereas with other models the values are nearly 4–11% higher. In a comparison between CD and CI fuzzy granulation based models (Table 4), the PA values for models 4 and 5 (CD models) are superior with the improvement of the accuracy by nearly 5–6% compared to models 2 and 3 (CI models), respectively. This clearly indicates that CD granules efficiently explored the class-wise belongingness of features to classes and provided an improved class discrimination information responsible for enhanced classification accuracy.

In a performance comparison of models with NRS and PaRS, it is observed from Table 4 that the PA values for model 5 (83.75) compared to model 4 (81.03), and model 3 (77.87) compared to model 2 (76.01) are higher. This specifies that the NRS based feature selection method restores better local information from neighborhood granules that is helpful for improved performance. Thus comparatively among the five models of classification, the model (model 5) that explored and incorporated granular feature space, CD fuzzy granulation and NRS based feature selection methods provided the best performance for all percentages of the training sets as observed in Table 4. Further, it is seen that the impact of CD fuzzy granule generation is more compared to NRS based feature selection in the classification performance. That is, in an environment of PaRS or NRS based feature selection methods, classification models with CD fuzzy granulation provided the PA value increment of about 5–6% (model 4 over model 2 and model 5 over model 3) over CI. In an environment of CD or CI fuzzy granulation based methods, classification models with NRS based feature selection performed well than PaRS based method with an increased PA value of about 2% (model 3 over model 2 and model 5 over model 4). Further, the performance with the combining effect of CI fuzzy granulation and NRS (model 3) is not comparable to CD fuzzy granulation and PaRS (model 4). These comparisons clearly justified the efficacy of CD granulation. The superiority of the proposed model to others is also validated with the kappa coefficient (KC) measure as shown in Table 4. All the critically assessed improved performance obtained with the PA is justified and supports the superiority claim of the proposed rough-fuzzy granulation and feature selection model with KC measure.

Table 4 also reveals that the accuracy obtained with the proposed model (model 5) for minimum percentage of training set is higher compared to the model incorporating CI fuzzy granulation and both PaRS and NRS based feature selection methods at 50% training set. This is particularly important when there is a scarcity of training set (e.g., land covers classification of real remote sensing images). This critically assessed improved performance claim is valid for both 20% and 50% training sets and is depicted in Table 4.

A comparative analysis in terms of total computational time T_c (sum of the training and testing times), as required by different models using k-NN classifier ($k=1$) for all three percentages of training sets, is depicted in Table 4. All the simulations are done in MATLAB (Matrix Laboratory) environment in Pentium-IV machine with 3.19 GHz processor speed. It is seen for all the cases that the T_c values (in s) for model 5, as expected, are higher compared to others with the cost of improved performance.

In addition to the above comparison of performance with PA, measures like Precision and Recall with the VOWEL data at 20% training set have been calculated for different models. We have selected models 1, 2 and 5 for comparison (Table 5) because the first one is based on the non-granulated feature space, second one with CI fuzzy granulation and PaRS, and third one with the best combination of CD fuzzy granulation and NRS. Thus a feel of comparison with the models that provided different combinations of results can be obtained. Although the measurements are made for all percentages of training sets, we have shown the results (Table 5) for 20% training set only because the claim for improvement with the proposed model is similar for all training sets. It is observed from the table that for all the classes and with both accuracy measurements the proposed model performed better than others in a class-wise agreement comparison.

Further, the performance comparison of models using β and Davies–Bouldin (DB) indexes are made with VOWEL data set. Here, the classifiers are trained with 20% of the data set and then

Table 5
Performance comparison in terms of Precision and Recall with the models 1, 2, and 5 using k-NN classifier ($k=1$), for all classes of VOWEL data at 20% training set ($\mathbf{p}=2$, $\Phi=0.45$).

Class	Model 1		Model 2		Model 5	
	Precision	Recall	Precision	Recall	Precision	Recall
1	66.05	75.28	69.43	78.77	75.68	92.71
2	74.67	77.80	76.62	82.13	83.31	89.56
3	83.08	89.29	86.03	92.79	90.39	96.75
4	80.02	87.86	84.65	93.44	93.39	95.75
5	68.44	78.21	70.89	80.77	76.64	91.01
6	78.93	86.21	82.12	92.23	86.76	93.79

Table 6
Performance comparison with dispersion score for all classes of VOWEL data at 20% training set ($\mathbf{p}=2$, $\Phi=0.45$).

Class	Model 1	Model 2	Model 3	Model 4	Model 5
1	0.9013	0.8298	0.8056	0.7846	0.7059
2	0.7134	0.5185	0.6295	0.4398	0.4248
3	0.4325	0.3467	0.2620	0.2601	0.2534
4	0.6132	0.5346	0.4386	0.4321	0.4422
5	0.7343	0.5338	0.5636	0.4979	0.4808
6	0.4262	0.3015	0.2021	0.2135	0.1950

Table 7
Performance comparison of models in terms of β and DB values using k-NN classifier ($k=1$) for VOWEL data ($\mathbf{p}=2$, $\Phi=0.45$).

Model	β	DB
Training samples	6.5421	0.7213
1	4.8463	1.3231
2	4.9901	1.2847
3	5.1033	1.2603
4	5.5664	1.1565
5	5.7352	1.1264

the said trained classifiers are applied on the whole data to partition into six categories. Results in terms of β and DB values are depicted in Table 7, which reveal the superiority of model 5 to others with respect to all aspects (i.e., (i) granulated over non-granulated feature space, (ii) CD over CI fuzzy granulation, and (iii) Pawlak's and neighborhood rough sets (NRS) based feature selection).

Dispersion measure: *index of class overlapping and performance evaluation*. Classifiers' efficiencies are also tested in terms of the dispersion score (DS) defined in Eq. (7). Using DS the dispersion of classified patterns into various classes is estimated and its physical significance with VOWEL data set is highlighted. Also we have analyzed and justified the superiority of the proposed model with this measure. Comparative results of all the classes with five classification models (with k-NN classifier ($k=1$)) for 20% training set using DS are depicted in Table 6.

Let us consider the results of confusion matrices and corresponding DS values obtained with models 3 and 5 for VOWEL data at 20% training set, as shown in Fig. 12. It is observed from the scatter plot of VOWEL data set (Fig. 11) that class 1 (/∂/) is highly overlapping with classes 5 (/e/) and 2 (/a/), and a small overlapping with class 6 (/o/). This is exactly reflected by the confusion matrix of model 5 (Fig. 12(b)), whereas for model 3 (Fig. 12(a)) this is not so. Accordingly the DS value is lower for model 5, signifying its superiority. The superiority of model 5–3

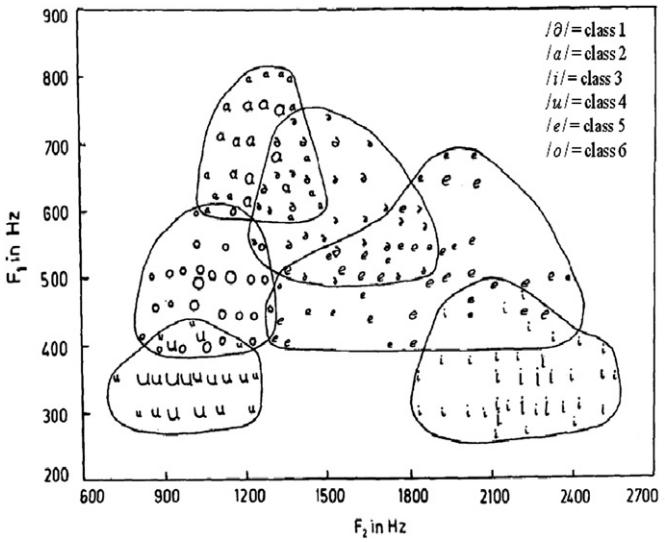

Fig. 11. Scatter plot of VOWEL data in F_1-F_2 plane.

and other models is similarly observed for all classes except class 4 (/u/), where DS value of model 5 is higher than those of models 3 and 4 (Table 6). Therefore, DS value may be viewed to provide an appropriate quantitative index in evaluating the overlapping of classes in terms of dispersion of misclassified patterns and to quantify the class-wise performance of classifiers accordingly.

Performance comparison of rough-fuzzy granulation and feature selection models using other classifiers. So far we have described the effectiveness of the proposed rough-fuzzy granulation and feature selection model using k-NN ($k=1$) classifier. In this section we describe the effectiveness of the same model using some other classifiers, e.g., k-NN ($k=3$ and 5), maximum likelihood (ML) classifier and multi-layered perceptron (MLP).

The comparative results of all models with these classifiers are depicted in Table 8 for training set of 20%, as an example. The superiority of model 5 to others for different sets of classifiers is evident. Also similar improvement in performance of the models (using different classifiers) with granulated over non-granulated, CD granulation over CI granulation and NRS based feature selection over PaRS is observed as in the case of k-NN ($k=1$) classifier.

Statistical significance test of the models using χ^2 test. To strengthen the claim of effectiveness of the proposed rough-fuzzy model for classification, we performed the statistical significance test using χ^2. The comparative results for all models are depicted in Table 9 for test set of 80%, as an example. We have selected models 1, 2 and 5 for comparison (Table 9), as we did for the evaluation in terms of Precision and Recall measures. It is observed from Table 9 that the p value obtained for model 5 is in the range of $0.30 < p < 0.50$, whereas it is $0.10 < p < 0.20$ and $0.05 < p < 0.10$, respectively for models 2 and 1. This observation shows that the significance of model 5 compared to models 1 and 2 is much higher and therefore justifies its superiority.

Performance comparison of rough-fuzzy granulation and feature selection models with fuzzy granulation and principal component analysis based feature reduction models. Until now we have compared the performance of fuzzy granulation and rough set based feature selection models. Comparison of these models is further made with fuzzy granulation and principal component analysis (PCA) based feature reduction models. We have considered two models based on CI and CD fuzzy granulation with PCA feature reduction. The results are shown in Table 10 for training set of 20% with k-NN classifier, as an example. Effectiveness of models with rough set feature selection compared to PCA is apparent.

a

		Predicted Class						Dispersion Score
	Class	C_1	C_2	C_3	C_4	C_5	C_6	
Actual Class	C_1	35	10	0	0	15	4	0.7059
	C_2	8	66	0	0	1	5	0.4248
	C_3	0	0	134	0	20	0	0.2534
	C_4	5	0	0	109	1	20	0.4422
	C_5	9	1	23	0	151	2	0.4908
	C_6	0	0	0	8	1	153	0.1950

b

		Predicted Class						Dispersion Score
	Class	C_1	C_2	C_3	C_4	C_5	C_6	
Actual Class	C_1	29	7	0	1	20	7	0.8056
	C_2	10	56	1	0	3	10	0.6295
	C_3	0	0	133	0	21	0	0.2620
	C_4	2	0	0	102	0	31	0.4386
	C_5	13	0	29	1	140	3	0.5636
	C_6	0	0	0	14	2	146	0.2621

Fig. 12. Analysis of dispersion measure with confusion matrix for VOWEL data with (a) model 3 and (b) model 5.

Table 8
Classification accuracies (PA) of models with different classifiers at 20% training set of VOWEL data ($\mathbf{p}=2$, $\Phi=0.45$).

Model	k-NN ($k=3$)	k-NN ($k=5$)	ML	MLP
1	74.20	73.63	75.21	76.33
2	76.11	76.34	77.34	78.06
3	77.78	77.89	78.34	79.87
4	81.03	80.86	82.25	83.75
5	83.91	84.01	83.87	84.88

Table 9
Statistical significance test of models at 80% test set with VOWEL data.

Model	p value
1	$0.10 > p > 0.05$
2	$0.20 > p > 0.10$
5	$0.50 > p > 0.30$

Table 10
Performance comparison of models using k-NN classifier ($k=1$) with VOWEL data for 20% training set.

Model	PA	KC
1	75.150	0.7212
2	77.010	0.7704
3	78.810	0.7853
4	81.370	0.8165
5	83.960	0.8253
CI granulation + PCA	76.876	0.7514
CD granulation + PCA	80.113	0.7981

Table 11
Performance comparison of models using k-NN classifier ($k=1$) with SATIMAGE data.

Model	10% of training set		20% of training set		50% of training set	
	PA	KC	PA	KC	PA	KC
1	73.070	0.6622	75.150	0.7112	77.230	0.7157
2	77.050	0.6985	78.060	0.7346	79.270	0.7384
3	78.820	0.7095	79.650	0.7401	80.760	0.7413
4	81.670	0.7815	81.980	0.7867	82.070	0.7889
5	83.14	0.7909	83.680	0.7934	83.950	0.7988

5.1.2. SATIMAGE data

Comparative analysis of the five classification models using k-NN classifier ($k=1$) for this data is presented in Table 11 for different training sets and for measures like percentage of accuracy (PA) and kappa coefficient (KC). It is seen that for all the training sets, the models with fuzzy granulated feature space provided greater PA values compared to model 1 (i.e., with non-granulated feature space); justifying the use of granular computing based methods for improving the performance. The proposed rough-fuzzy model (model 5) is seen to be more effective with CD fuzzy granulation, and NRS based feature selection methods. For example, see the improvement of model 5 over 3 and model 4 over 2. As in the case of VOWEL data, model 5 (i.e., combination of CD granulation and NRS based feature selection) performed the best. Other findings and issues as discussed in the case of VOWEL data, e.g., effect of changing $\mathbf{p}$ and Φ values, significance of DS measure, performance with the measures like Precision and Recall, performance of models using other classifiers, are also found to be true here for the SATIMAGE data.

Table 12
Performance comparison of models using k-NN classifier ($k=1$) with WAVEFORM data.

Model	10% of training set		20% of training set		50% of training set	
	PA	KC	PA	KC	PA	KC
1	72.890	0.6156	74.460	0.6302	76.950	0.6677
2	77.178	0.6987	78.130	0.7105	78.970	0.7122
3	79.302	0.7011	79.840	0.7112	80.020	0.7136
4	81.460	0.7489	81.980	0.7511	82.110	0.7523
5	83.870	0.7526	84.210	0.7554	84.950	0.7581

5.1.3. WAVEFORM data

Performance comparison of models using k-NN classifier ($k=1$) for three percentages of training sets with this data set is made and results are depicted in Table 12. It is observed from Table 12 that for all training sets, model 5 yields superior results compared to models 1, 2, 3 and 4, in terms percentage of accuracy (PA) and kappa coefficient (KC). The superiority of model 4–2 and model 5–3 justifies the advantages of using CD fuzzy granulation. In another observation, superiority of model 5–4 and model 3–2 validates the effectiveness of NRS based feature selection. Similar to VOWEL data, all the analyses hold true for WAVEFORM data.

5.1.4. CALDONAZZO data

The performance comparison of results in terms of PA and KC measures with this data set is shown in Table 13. It is seen that with 10%, 20% and 50% training sets, the performance of the proposed model (model 5) is encouraging compared to rest four models. All critical assessments (as performed with VOWEL data) with CALDONAZZO data are supporting the improved performance of the proposed model that explores the mutual advantages of CD fuzzy granulated feature space and NRS based feature selection.

5.1.5. PHONEME data

It is observed from Table 14 that the proposed rough-fuzzy model provides a higher classification accuracy compared to those obtained from models with non-granulated feature space,

Table 13
Performance comparison of models using k-NN classifier ($k=1$) with CALDONAZZO data.

Model	10% of training set		20% of training set		50% of training set	
	PA	KC	PA	KC	PA	KC
1	72.550	0.5402	73.780	0.5699	77.050	0.6172
2	74.010	0.6013	76.160	0.6112	78.180	0.6679
3	77.710	0.6115	78.653	0.6226	79.910	0.6713
4	80.650	0.6501	81.140	0.6689	82.450	0.7201
5	82.180	0.6653	82.710	0.6815	83.910	0.7315

Table 14
Performance comparison of models using k-NN classifier ($k=1$) with PHONEME data.

Model	10% of training set		20% of training set		50% of training set	
	PA	KC	PA	KC	PA	KC
1	76.45	0.4502	79.12	0.4682	80.02	0.5001
2	78.37	0.4983	79.76	0.5210	80.11	0.5623
3	80.01	0.5420	80.54	0.5632	80.98	0.5796
4	81.72	0.5889	82.03	0.5980	82.37	0.6103
5	83.03	0.5974	83.89	0.6137	84.11	0.6572

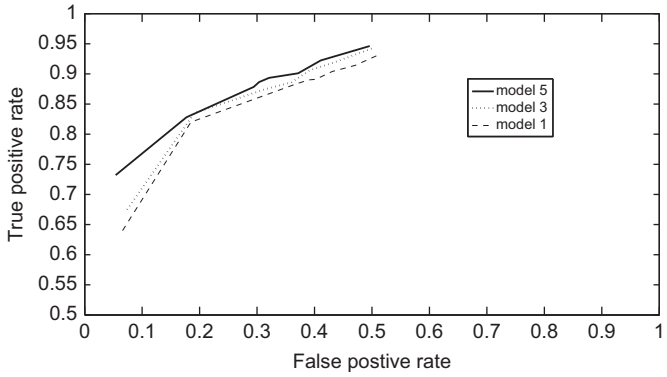

Fig. 13. ROC plot of models 1, 3, and 5 with PHONEME data set.

CI granulation with Pawlak's rough set and NRS feature selection, and CD granulation with Pawlak's rough set feature selection. This is true for all three sets of training data. Further, the superiority of the proposed model to others with all types of analyses are also achieved for this data sets, as in the case of VOWEL data. Further, we have compared the models' performance in terms of ROC curve and is plotted in Fig. 13 for the models 1, 3 and 5. Fig. 13 clearly reveals the superiority of model 5 to others.

5.1.6. PAGE-BLOCK data

Like different data sets (Table 3), the superiority of model 5 to others in terms of PA and KC are observed with PAGE-BLOCK data, as shown in Table 15. Various observations that justify the advantages in the use of proposed rough-fuzzy model (model 5) are also found to be similar with this data set, as in the case of remaining data sets.

5.1.7. Synthetic image

Noisy synthetic remote sensing images with different σ (Fig. 8(b)) values are used to compare the performance of five models using k-NN classifier ($k=1$), in terms of percentage of accuracy (PA) and the corresponding results are shown in Table 16 for 20% training set.

Table 15
Performance comparison of models using k-NN classifier ($k=1$) with PAGE-BLOCK data.

Model	10% of training set		20% of training set		50% of training set	
	PA	KC	PA	KC	PA	KC
1	86.02	0.5880	86.55	0.5708	88.90	0.6031
2	88.39	0.6122	88.88	0.6243	90.03	0.6473
3	90.76	0.6436	91.01	0.6532	92.22	0.6780
4	91.85	0.6532	92.73	0.6679	93.15	0.7011
5	93.89	0.6691	94.65	0.7101	95.89	0.7382

Table 16
Classification accuracies (PA) of models using k-NN classifier ($k=1$) for synthetic image at 20% training set (different σ) ($\mathbf{p}=2$, $\Phi=0.45$).

Model	PA			
	$\sigma=1$	$\sigma=2$	$\sigma=3$	$\sigma=4$
1	95.72	82.03	72.17	60.77
2	97.86	91.10	77.23	63.01
3	98.12	92.23	80.23	67.83
4	98.83	94.34	82.36	69.36
5	99.54	95.87	84.41	71.44

The table revealed the superiority of the proposed model (model 5) to others for all the noise levels. Since similar trend of observation, as discussed in the case of VOWEL data, is obtained with other measures for the synthetic remote sensing image, we have not put those results here. Fig. 14 shows the resulting classified images obtained by models 1 and 5 for the noisy input image with $\sigma=2$ (i.e., Fig. 8(b)). Superiority of model 5–1, as indicated in Table 16, is further verified visually from Fig. 14. Here we have shown the classified images obtained from these two models, as an example, because one of them performed the worst and the other performed the best.

5.2. Classification of partially labeled data sets

In Section 5.1 we have demonstrated the performance of the proposed model for classification of completely labeled data sets. In this section we describe the same on two partially labeled data, namely IRS-1A and SPOT images. Here the classifiers are initially trained with labeled data of six land cover types and then the said trained classifiers are applied on the unlabeled image data to partition into six regions.

5.2.1. IRS-1A image

IRS-1A image is classified with five different models using k-NN classifier ($k=1$), and the performance comparison in terms of β value and Davies–Bouldin (DB) value is shown in Table 17. As expected, the β value is the highest and DB value is the lowest for the training set (Table 17). It is also seen that the proposed model yielded superior results in terms of both the indexes, compared to other four models. As a whole the gradation of performance of five models can be established with the following β relation:

$$\beta_{training} > \beta_{proposed} > \beta_{model4} > \beta_{model3} > \beta_{model2} > \beta_{model1}. \quad (11)$$

Similar gradation of performance is also observed with DB values, which further supports the superiority of the proposed model.

In order to demonstrate the significance of granular computing visually, let us consider Fig. 15(a) and (b) depicting the output corresponding to models 1 (without granulation) and 5 (with granulation), say. It is clear from the figures that the proposed model 5 performed well in segregating different areas by properly classifying the land covers. For example, the *Howrah bridge* over

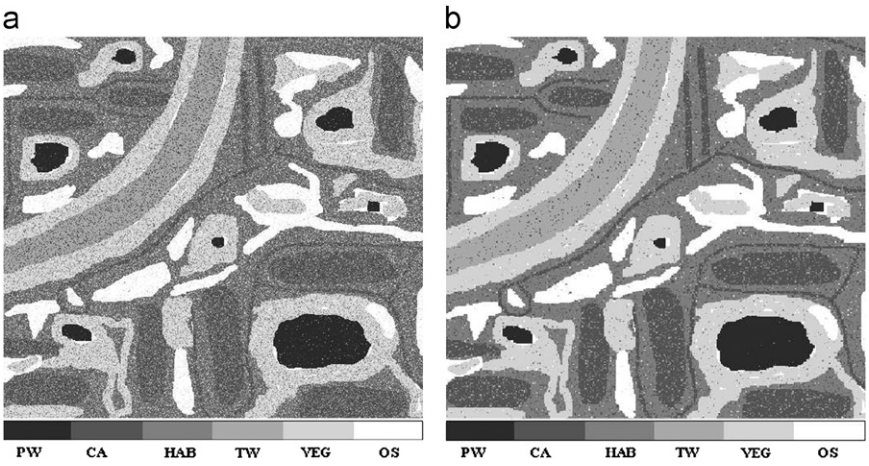

Fig. 14. Classified synthetic image (for $\sigma = 2$) by (a) model 1 and (b) model 5 (proposed model).

Table 17
Performance comparison of models using k-NN classifier ($k=1$) with partially labeled data sets ($\mathbf{p}=2$, $\Phi = 0.45$).

Model	β value		DB value	
	IRS-1A	SPOT	IRS-1A	SPOT
Training samples	9.4212	9.3343	0.5571	1.4893
1	6.8602	6.8745	0.9546	3.5146
2	7.1343	7.2301	0.9126	3.3413
3	7.3559	7.3407	0.8731	3.2078
4	8.1372	8.2166	0.7790	2.8897
5	8.4162	8.4715	0.7345	2.7338

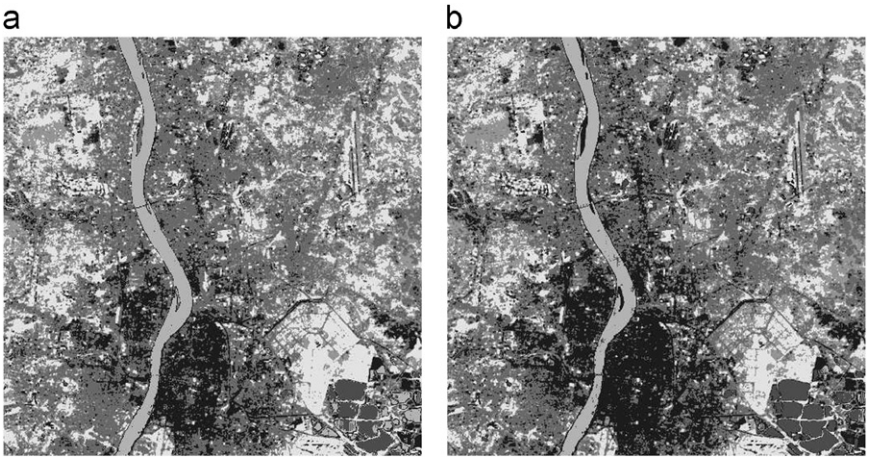

Fig. 15. Classified IRS-1A images with (a) model 1 and (b) model 5 (proposed model).

the south part of the *river* is more prominent in Fig. 15(b), whereas it is not so in Fig. 15(a). A zoomed version of the said *bridge* region is shown in Fig. 16(a) and (b) to have an improved visualization. Similarly, the regions such as *Saltlake stadium* and *water bodies* are more distinct and well shaped with model 5 as shown in Fig. 16(d) (zoomed version). These observations further justify the significance of the β and DB indexes in reflecting the performance of the models automatically without visual intervention.

5.2.2. SPOT image

With SPOT image, the comparative results of five models using k-NN classifier ($k=1$) in terms of β and DB values are shown in Table 17, which revealed the supremacy of the proposed model (model 5). The significance of model 5 is further justified visually from Fig. 17 that illustrates the classified images corresponding to models 1 and 5. Fig. 17(b) is superior in the sense that the different structures (e.g., *roads* and *canals*) are more prominent.

6. Conclusions

In the present article, we described a rough-fuzzy model for pattern classification. The model formulates a class-dependent (CD) fuzzy granulation of input feature space, where the membership functions explore the degree of belonging of features into

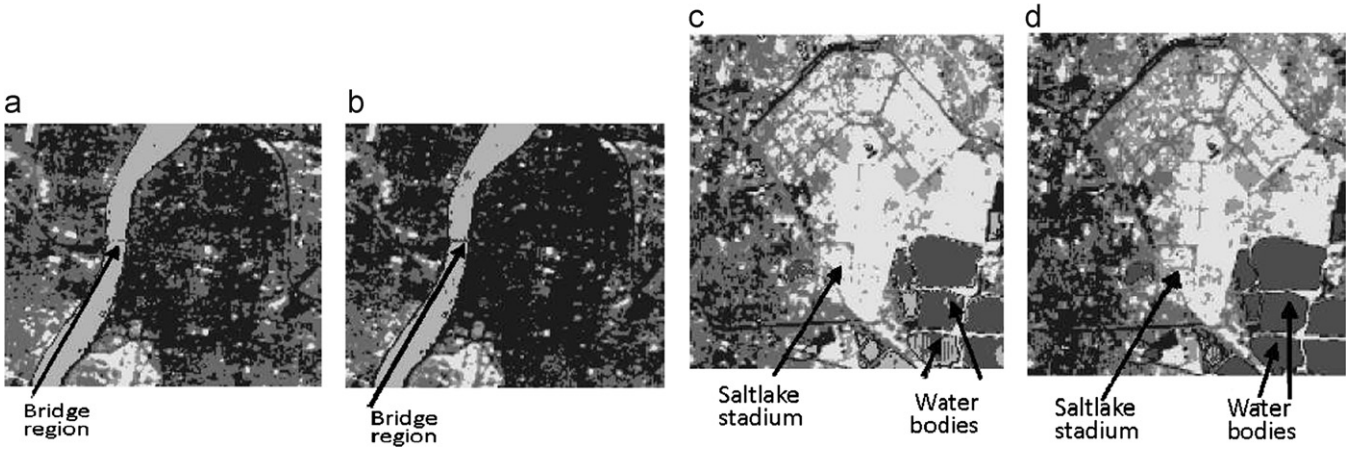

Fig. 16. (Zoomed) Two selected regions of classified IRS-1A image with (a and c) model 1 and (b and d) model 5.

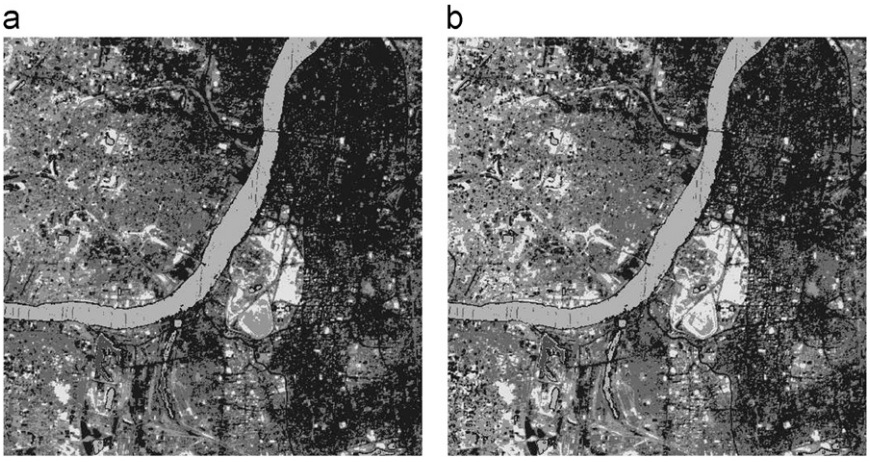

Fig. 17. Classified SPOT images with (a) model 1 and (b) model 5 (proposed model).

different classes and make it more suitable for improved class label estimation. The advantage of neighborhood rough sets that deal with both numerical and categorical data without any discretisation is also realized in the proposed model. The neighboring concept facilitates to gather the local/contextual information through neighbor granules that provide improved class discrimination information. We have defined a dispersion measure for classifiers' performance that reflects well the overlapping characteristics of a class with others and can be viewed as an appropriate index in evaluating the class-wise performance of a classifier. It may be mentioned here that fuzzy granulation of feature space described in [13] for case generation is similar to the method of class-independent granulation used here.

With extensive experimental results on various types of real life as well synthetic data, both fully and partially labeled, it is found that the effect of CD fuzzy granulation is substantial compared to the rough feature selection methods in improving classification performance and the combined effect is further encouraging for the data sets with highly overlapping classes. The statistical significance of the proposed model is also supported by the χ^2 test.

The computational complexity of the proposed model is little high. However, its learning ability with small percentage of training samples will make it practicably applicable to problems with a large number of overlapping classes. While the classification accuracy appears to drop drastically after $\Phi = 0.5$, it is interesting and also beneficial to note that highest classification accuracy is maintained for a moderately wide range of Φ for all types of distances.

Acknowledgments

Two of the authors (Sankar K. Pal and Saroj K. Meher) acknowledge the Center for Soft Computing Research: A National Facility, funded by the Department of Science and Technology, Govt. of India. The work was done while Prof. Pal held J.C. Bose Fellowship of the Govt. of India.

References

[1] A. Bargiela, W. Pedrycz, Granular Computing: An Introduction, Kluwer Academic Publishers, Boston, 2003.
[2] A. Skowron, J.F. Peters, Rough-granular computing, in: W. Pedrycz, A. Skowron, V. Kreinovich (Eds.), Handbook of Granular Computing, John Wiley & Sons Ltd., West Sussex, England, 2008, pp. 285–328.
[3] W. Pedrycz, B.J. Park, S.K. Oh, The design of granular classifiers: a study in the synergy of interval calculus and fuzzy sets in pattern recognition, Pattern Recognition 41 (2008) 3720–3735.
[4] L.A. Zadeh, Fuzzy sets and information granularity, in: M. Gupta, R. Ragade, R. Yager (Eds.), Advances in Fuzzy Set Theory and Applications, North-Holland Publishing Co., Amsterdam, 1979, pp. 3–18.
[5] Z. Pawlak, Rough sets, International Journal of Computer and Information Science 11 (1982) 341–356.
[6] J.F. Peters, Z. Pawlak, A. Skowron, A rough set approach to measuring information granules, in: Proceedings of Annual International Conference on Computer Software and Applications, 2002, pp. 1355–1360.

[7] L.A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems 90 (1997) 111–127.
[8] L.A. Zadeh, Toward a generalized theory of uncertainty (GTU)—an outline, Information Sciences 172 (2005) 1–40.
[9] L.A. Zadeh, Is there a need for fuzzy logic? Information Sciences 178 (2008) 2751–2779.
[10] S.K. Pal, P. Mitra, Pattern Recognition Algorithms for Data Mining, CRC Press, Boca Raton, FL, 2004.
[11] S.K. Pal, A. Skowron (Eds.), Rough-Fuzzy Hybridization: A New Trend in Decision Making, Springer-Verlag, Singapore, 1999.
[12] M. Banerjee, S. Mitra, S.K. Pal, Rough fuzzy MLP: knowledge encoding and classification, IEEE Transactions on Neural Networks 9 (1998) 1203–1216.
[13] S.K. Pal, P. Mitra, Case generation using rough sets with fuzzy representation, IEEE Transactions on Knowledge and Data Engineering 16 (2004) 293–300.
[14] S.K. Pal, S. Mitra, Multilayer perceptron, fuzzy sets, and classification, IEEE Transactions on Neural Networks 3 (1992) 683–697.
[15] Z. Pawlak, Rough Sets—Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Netherlands, 1991.
[16] T.Y. Lin, Granulation and nearest neighborhoods: rough set approach, in: W. Pedrycz (Ed.), Granular Computing: An Emerging Paradigm, Physica-Verlag, Heidelberg, Germany, 2001, pp. 125–142.
[17] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Information Sciences 178 (2008) 3577–3594.
[18] Q. Hu, Z. Xie, D. Yu, Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognition 40 (2007) 3509–3521.
[19] Q. Hu, J. Liu, D. Yu, Mixed feature selection based on granulation and approximation, Knowledge-Based Systems 21 (2008) 294–304.
[20] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley-Interscience Publications, USA, 2000.
[21] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed., Prentice Hall, NJ, USA, 1998.
[22] J. Han, M. Kamber, Data Mining: Concepts and Techniques, 2nd ed., Morgan Kaufmann, San Francisco, CA, 2006.
[23] J. Cohen, A coefficient of agreement for nominal scale, Education and Psychological Measurement 20 (1960) 37–46.
[24] S.K. Pal, A. Ghosh, B. Uma Shankar, Segmentation of remotely sensed images with fuzzy thresholding, and quantitative evaluation, International Journal of Remote Sensing 21 (2000) 2269–2300.
[25] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (1979) 224–227.
[26] L.A. Zadeh, Fuzzy sets, Information and Control 8 (1965) 338–353.
[27] S.K. Pal, D.D. Majumder, Fuzzy Mathematical Approach to Pattern Recognition, Wiley (Halsted), NY, USA, 1986.
[28] A. Chouchoulas, Q. Shen, Rough set-aided keyword reduction for text categorisation, Applied Artificial Intelligence 15 (2001) 843–873.
[29] R.G. Congalton, R. Mead, A quantitative method to test for consistency and correctness in photointerpretation, Photogrammetric Engineering and Remote Sensing 49 (1983) 69–74.
[30] S.K. Pal, D.D. Majumder, Fuzzy sets and decision making approaches in vowel and speaker recognition, IEEE Transactions on Systems, Man, and Cybernetics 7 (1977) 625–629.
[31] Elena Database, 〈ftp://ftp.dice.ucl.ac.be/pub/neural-nets/ELENA/databases/REAL/〉.
[32] A. Asuncion, D. Newman, UCI Machine Learning Repository, 2007.
[33] NRSA, IRS Data Users Hand Book, Technical Report, 1989. Document No. IRS/NRSA/NDC/HB-02/89.
[34] D.P. Mandal, C.A. Murthy, S.K. Pal, Formulation of a multivalued recognition system, IEEE Transactions on Systems, Man, and Cybernetics 22 (1992) 607–620.
[35] F. Melgani, B.A.R. Al Hashemy, S.M.R. Taha, An explicit fuzzy supervised classification method for multispectral remote sensing images, IEEE Transactions on Geoscience and Remote Sensing 38 (2000) 287–295.
[36] S.K. Meher, B. Uma Shankar, A. Ghosh, Wavelet-feature-based classifiers for multispectral remote-sensing images, IEEE Transactions on Geoscience and Remote Sensing 45 (2007) 1881–1886.
[37] J. Fu, H.J. Caulfield, S.-M. Yoo, D. Wu, Fuzzy aggregation with artificial color filters, Information Science 180 (2010) 167–180.

Sankar K. Pal (www.isical.ac.in/~sankar) is a *Distinguished Scientist* of the Indian Statistical Institute and a former Director. He is also a J.C. Bose Fellow of the Govt. of India. He founded the Machine Intelligence Unit and the Center for Soft Computing Research: A National Facility in the Institute in Calcutta. He received a Ph.D. in Radio Physics and Electronics from the University of Calcutta in 1979, and another Ph.D. in Electrical Engineering along with DIC from Imperial College, University of London in 1982. He joined his Institute in 1975 as a CSIR Senior Research Fellow where he later became a Full Professor in 1987, Distinguished Scientist in 1998 and the Director for the term 2005–2010.

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 in US Naval Research Laboratory, Washington, DC in 2004. Since 1997 he has been serving as a *Distinguished Visitor* of IEEE Computer Society (USA) for the Asia-Pacific Region, and held several visiting positions in Italy, Poland, Hong Kong and Australian universities.

Prof. Pal is a *Fellow* of the IEEE, USA, the Academy of Sciences for the Developing World (TWAS), Italy, International Association for Pattern recognition, USA, International Association of Fuzzy Systems, USA, and all the four National Academies for Science/Engineering in India. He is a co-author of 17 books and more than 400 research publications in the areas of Pattern Recognition and Machine Learning, Image Processing, Data Mining and Web Intelligence, Soft Computing, Neural Nets, Genetic Algorithms, Fuzzy Sets, Rough Sets and Bioinformatics.

He has received the 1990 S.S. *Bhatnagar Prize* (which is the most coveted award for a scientist in India), 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 (USA), 1994 IEEE Trans. Neural Networks Outstanding Paper Award (USA), 1995 NASA Patent Application Award (USA), 1997 IETE-R.L. Wadhwa Gold Medal, the 2001 INSA-S.H. Zaheer Medal, 2005–2006 Indian Science Congress-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.

Prof. Pal is/was an *Associate Editor* of IEEE Trans. Pattern Analysis and Machine Intelligence (2002–2006), IEEE Trans. Neural Networks (1994–1998 and 2003–2006), Neurocomputing (1995–2005), Pattern Recognition Letters (1993–2011), Int. J. Pattern Recognition & Artificial Intelligence, Applied Intelligence, Information Sciences, Fuzzy Sets and Systems, Fundamenta Informaticae, LNCS Trans. On Rough Sets, Int. J. Computational Intelligence and Applications, IET Image Processing, J. Intelligent Information Systems, and Proc. INSA-A; *Editor-in-Chief*, Int. J. Signal Processing, Image Processing and Pattern Recognition; a Book Series Editor, Frontiers in Artificial Intelligence and Applications, IOS Press, and Statistical Science and Interdisciplinary Research, World Scientific; a *Member*, *Executive Advisory Editorial Board*, IEEE Trans. Fuzzy Systems, Int. Journal on Image and Graphics, Int. J. Computational Science & Engineering, and Int. J. Approximate Reasoning; and *Guest Editor*, IEEE Computer, and Theoretical Computer Science—C.

Saroj K. Meher is Faculty of Systems Science and Informatics Unit, Indian Statistical Institute, India. He received the Ph.D. degree from National Institute of Technology, Rourkela, India, in 2003. His research interest includes granular computing, pattern recognition, soft computing methods, and digital signal processing. He has published many research articles in internationally reputed journals and refereed conferences.

Soumitra Dutta is an authority on all aspects of innovation in the knowledge economy, with a refreshing global perspective. Throughout his distinguished career, he has focussed on how to drive business growth through the right combination of innovative people and technology. This is particularly relevant in the current crisis when innovation is the best answer for both thriving today and emerging stronger in a post-crisis world. He has co-written several important books on technology-enabled business innovation. In *Innovating at the Top*, he offers proven ways that senior executives can improve innovation performance distilled from interviews with the CEOs of nine highly innovative international corporations. He has also co-edited eight annual reports for the World Economic Forum on the impact of information technology on development and national competitiveness. Dr. Dutta is the Roland Berger Professor of Business and Technology and Founder and Academic Director of ELab at INSEAD.