Title: Hyperspectral Band Selection Based on Rough Set

This paper appear in: IEEE Transactions on Geoscience and Remote Sensing

Date of Publication: 2015

Author(s): Swarnajyoti Patra, Prahlad Modi, and Lorenzo Bruzzone

Volume: 53, Issue: 10

Page(s): 5495-5503

DOI: 10.1109/TGRS.2015.2424236

Hyperspectral Band Selection Based on Rough Set

Swarnajyoti Patra, Prahlad Modi, and Lorenzo Bruzzone

***Abstract*—Band selection is a well known approach to reduce the dimensionality of hyperspectral imagery. Rough set theory is a paradigm to deal with uncertainty, vagueness, and incompleteness of data. Although it has been applied successfully to feature selection in different application domains, it is seldom used for the analysis of the hyperspectral imagery. In this paper, a Rough set based supervised method is proposed to select informative bands from hyperspectral imagery. The proposed technique exploits Rough set theory to compute the relevance and significance of each spectral band. Then by defining a novel criterion it select the informative bands that have higher relevance and significance values. To assess the effectiveness of the proposed band selection technique, three state-of-the-art methods (one supervised and two unsupervised) used in the remote sensing literature are analyzed for comparison on three hyperspectral data sets. The results of this comparison points to the superiority of the proposed technique, especially when a small number of bands is to selected.**

***Index Terms*—Feature selection, feature extraction, Rough sets, support vector machine, remote sensing, hyperspectral imagery.**

I. Introduction

Hyperspectral images are characterized by hundreds of bands acquired in a contiguous spectral range and narrow spectrum interval. A hyperspectral image can be viewed as an image cube where the first two dimensions indicate the spatial coordinate of the image and the third represents the number of bands of the image [1], [2]. Thus, each pixel represents a pattern whose number of attributes is equal to the number of bands. Due to the availability of a large number of bands, "the curse of dimensionality" and computation complexity are become two critical issues for the processing of hyperspectral imagery. The "curse of dimensionality" can be avoided by providing a sufficiently large number of training samples. However, in many real applications this is not feasible. Moreover, due to the high dimensionality of hyperspectral imagery, the data volume to be processed is generally huge. As a result, the computational complexity for the analysis of hyperspectral imagery is very high. A simpler way to address such problems is to reduce the dimensionality of the hyperspectral data.

Hyperspectral images have a large number of highly correlated bands, thus containing redundant information [3]. Accordingly, removing appropriate bands may reduce this redundancy without decreasing of the useful information. There are two main approaches to reduce the number of band: one is feature/band selection [1], [4]–[19] (which consists in selecting some informative bands with low correlation among them); the other is feature/band extraction [20]–[26] (which compresses all the bands using mathematical transformation). The difference between the two approaches is substantial. In this paper we focus on feature (band) selection techniques, which preserve the original physical information of the acquired spectral channels. This has several advantages: 1) it makes it possible a conceptual validation of the selected features and thus of the information used by the classifier; 2) the results of feature selection can be used as a data mining tool for inferring the physical information (i.e., spectral channels) on the basis of which the classes are discriminated; and 3) the selection of a subset of original bands results in the possibility to define a system in which irrelevant features are not acquired and stored for the considered application. This can simplify the acquisition process (reducing the number of bands to be acquired) and reduce both the processing time and the data storage requirements.

Depending on the availability of labeled reference data, band selection methods are categorized into two groups i.e., supervised [1], [5]–[13] and unsupervised [1], [14]–[19]. Supervised methods need a training set (i.e., the class label information of a subset of patterns), whereas unsupervised methods do not assume the availability of labeled patterns. Unsupervised methods often evaluate the importance of a band in classification by using various statistical measures such as correlation, mutual information, or using clustering quality assessment. On the other hand, supervised methods exploit the labeled patterns for training. Thus, the selected bands usually provide higher classification accuracy than those selected by unsupervised techniques.

There are many supervised and unsupervised band selection methods presented in the literature for the analysis of hyperspectral imagery. Chang et al. [1] proposed two eigenanalysis-based criteria for band prioritization: i) an unsupervised PCA-based criterion, and ii) a supervised classification-based criterion. After band prioritization, they further proposed a divergence-based band decorrelation to remove the redundant bands. The supervised techniques based on the Jeffries-Matusita distance, divergence, and Bhattacharya distance between classes are widely used as band selection criteria in hyperspectral data [5], [8]–[10]. In [6] a suboptimal band subset is selected by minimizing an estimated error of the Bayes classifier. A Rough set and fuzzy C-Means based supervised band selection technique is presented in [11]. A spatially invariant supervised hyperspectral band selection technique is presented in [12]. In [13], Yang et al. presented a fast supervised band selection method based on the covariance matrix for hyperspectral image classification. In [14], four different unsupervised criteria are proposed for hyperspectral band selection. In [15], Du and Yang proposed an unsupervised similarity measure by using linear prediction and orthogonal subspace projection to determined informative bands. Martinez-Uso et al. [16] adopted a clustering technique to group similar bands into a cluster. Then the most informative bands are selected by applying either a mutual information criterion or a Kullback-Leibler (KL) divergence criterion. In

[17], an affinity propagation based clustering technique is presented to select appropriate bands.

Rough set theory is a paradigm to deal with uncertainty, vagueness, and incompleteness of data [27]–[29]. It has been applied successfully to feature selection of discrete valued data [30]–[32]. The quick reduct algorithm, the discernibility matrix based method, dynamic reducts, *etc.* are popular Rough set based feature selection methods [33]–[35]. Different heuristic approaches based on Rough set are also proposed in the literature [36], [37]. All these methods try to identify the most informative subset of features (also known as reduct) from the original feature set. To find the optimal reduct (or a solution close to the optimal reduct) they generate a large number of reducts that makes these techniques computationally demanding. Moreover, these methods detect the optimal reduct without considering the redundancy among the selected features. In [38], [39], Maji and Paul exploit Rough set theory to select the informative and nonredundant features of microarray gene expression data by avoiding the generation of large number of reducts.

Rough set theory has been applied successfully to feature selection in different application domains. It is seldom used for hyperspectral band selection. In this paper, a supervised hyperspectral band selection method is proposed inspired by the Rough set based feature selection technique presented in [38], [39]. The proposed technique exploits Rough sets theory to compute the relevance and significance of each band and then select informative bands having higher relevance and significance values. Moreover, a new criterion is presented to select most informative bands. The proposed technique uses simple first-order incremental search to avoid the generation of large numbers of reducts, thus making it less computationally demanding. The performance of the proposed approach was compared with three state-of-the-art methods (one supervised and two unsupervised) using the predictive accuracy of support vector machine on three different hyperspectral data sets. Experimental results show effectiveness of the proposed approach, especially when small number of bands should be selected.

The rest of this paper is organized as follows. Section II introduces the notions related to Rough sets. The proposed Rough set based band selection technique is presented in Section III. Section IV provides the description of the three hyperspectral data sets used for experiments. Section V presents different experimental results obtained on the considered data sets. Finally, Section VI draws the conclusion of this work.

II. Rough sets

A data set is described as a table where each row represents a pattern and each column represents an attribute that can be measured for each pattern. This table is called information system or approximation space [27]. In greater details, it is a pair (U, A), where $U = \{x_1, x_2, ..., x_n\}$ is a nonempty finite set of patterns called the universe, and A is a family of attributes such that $f : U \times A \rightarrow V$, where V is the value domain of A and f is an information function.

The theory of Rough sets exploits the notion of information system. Let (U, A) be an information system; then any subset P of A defines an equivalence relation $IND(P)$ as [27]:

$$IND(P) = \{(x_i, x_j) \in U \times U \mid \forall a \in P, f(x_i, a) = f(x_j, a)\} \tag{1}$$

$IND(P)$ is also called P-indiscernibility relation. If $(x_i, x_j) \in IND(P)$, then objects x_i and x_j are indiscernible from each other by attributes P. The partition of U generated by $IND(P)$ is denoted as [27]:

$$U/IND(P) = \{[x_i]_P \mid x_i \in U\} \tag{2}$$

where $[x_i]_P$ is the equivalence class of the P-indiscernibility relation containing x_i. The equivalence class of $IND(P)$ and the empty set ϕ are the elementary sets in the information system (U, A). Let (U, A) be a given information system and let $P \subseteq A$ and $X \subseteq U$. In general, it may not be possible to describe X precisely using only the information contained in P. One may characterize X by constructing the P-lower and P-upper approximations defined as follows [27]:

$$\begin{aligned} \underline{P}(X) &= \bigcup\{[x_i]_P \mid [x_i]_P \subseteq X\} \\ \overline{P}(X) &= \bigcup\{[x_i]_P \mid [x_i]_P \cap X \neq \phi\} \end{aligned} \tag{3}$$

The lower approximation $\underline{P}(X)$ is the union of all the elementary sets that are subsets of X, and the upper approximation $\overline{P}(X)$ is the union of all the elementary sets that have a nonempty intersection with X. Thus, the patterns in $\underline{P}(X)$ are definitely classified as a members of X on the basis of knowledge in P, while the patterns in $\overline{P}(X)$ are classified as possible members of X on the basis of knowledge in P. The lower approximation $\underline{P}(X)$ is also called positive region, denoted as $POS_P(X)$. The set $BN_P(X) = \overline{P}(X) - \underline{P}(X)$ is called the boundary region of X, and consists of those patterns that we cannot classify with high confidence into X on the basis of the knowledge in P. A set is said to be Rough if the boundary region is nonempty, otherwise it is crisp.

In many applications the class label of patterns are known. This a posteriori knowledge is represented by an attribute called decision attribute. The other attributes of patterns are called conditional attributes. An information system (U, A) is called a decision system if the attribute set $A = C \cup D$, where C and D represent the condition and decision attribute sets, respectively. The dependency between C and D can be defined as [27]:

$$\gamma_C(D) = \frac{\mid POS_C(D) \mid}{\mid U \mid} \tag{4}$$

Here $POS_C(D) = \bigcup \underline{C} X_i$, where $X_i = [x_i]_D$ i.e., X_i is the *i*th equivalence class induced by D and $\mid . \mid$ denotes the cardinality of a set.

In greater details, let (U, A) be a decision system as shown in table I, where $U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ and $A = C \cup D$ is a finite set of attributes. Here, conditional attributes C={Age, LEMS (Lower Extremity Motor Score)} and decision attribute D={Walk}. According to the indiscernibility relation defined in (1), the following partitions of U are created by different attributes of A:

$$U/IND(\{Age\}) = \{\{x_1, x_2, x_6\}, \{x_3, x_4\}, \{x_5, x_7\}\}$$

TABLE I
AN EXAMPLE OF DECISION TABLE

	Age	LEMS	Walk
x_1	16-30	50	yes
x_2	16-30	0	no
x_3	31-45	1-25	no
x_4	31-45	1-25	yes
x_5	46-60	26-49	no
x_6	16-30	26-49	yes
x_7	46-60	26-49	no

$$U/IND(\{LEMS\}) = \{\{x_1\},\{x_2\},\{x_3,x_4\},\{x_5,x_6,x_7\}\}$$
$$U/IND(\{Age,LEMS\}) = \{\{x_1\},\{x_2\},\{x_3,x_4\},\{x_5,x_7\},\{x_6\}\}$$
$$U/IND(\{Walk\}) = \{\{x_1,x_4,x_6\},\{x_2,x_3,x_5,x_7\}\}.$$

The positive region contains all objects of U that can be classified into classes of U/D using the knowledge in attributes C. Thus, for the above example, if $C = \{Age, LEMS\}$, the positive region is as follows:

$$POS_C(D) = \bigcup\{\{x_1\},\{x_2\},\phi,\{x_5,x_7\},\{x_6\}\} = \{x_1,x_2,x_5,x_6,x_7\}.$$

So the dependency between C and D is:

$$\gamma_C(D) = \frac{\mid POS_C(D) \mid}{\mid U \mid} = \frac{5}{7}.$$

The dependency measure is an important variable for finding out informative attributes P from C ($P \subseteq C$). If $\gamma_P(D) = 1$, D depends totally on P; if $0 < \gamma_P(D) < 1$, D depends partially on P; and if $\gamma_P(D) = 0$, then D does not depend on P.

The importance of an attribute to calculate the dependency on a decision attribute can be computed by measuring the significance of that attribute [38]. If we remove an attribute from a set of conditional attributes, the change in dependency is the measure of significance of that attribute. High changes in dependency indicate highly significant attributes. If there is no change in dependency then the attribute is not useful. The significance of an attribute $a \in C$ is computed as follows [38]:

$$\delta_C(D,a) = \gamma_C(D) - \gamma_{C-a}(D). \tag{5}$$

Considering the decision system of Table I, let $C = \{Age, LEMS\}$ and $D = \{Walk\}$. The significance of the attributes {Age} and {LEMS} is computed as follows [38]:

$$\delta_C(D,Age) = \gamma_C(D) - \gamma_{LEMS}(D) = \frac{5}{7} - \frac{2}{7} = \frac{3}{7}$$
$$\delta_C(D,LEMS) = \gamma_C(D) - \gamma_{Age}(D) = \frac{5}{7} - \frac{2}{7} = \frac{3}{7}$$

III. PROPOSED HYPERSPECTRAL BAND SELECTION METHOD

As mentioned before, hyperspectral images contain a large number of bands and many of them are redundant and/or weakly informative. The presence of such irrelevant bands may lead to both the "curse of dimensionality" and a computationally demanding classification task. Our main objective is to develop a band selection technique that can identify an effective subset of bands from the available hyperspectral data without decreasing the discrimination capability. Ideally, the selected bands should have high relevance with the classes, so that the prediction probability that a pattern belong to a specific classes will be high. However, if irrelevant bands are present in the selected subset, they may introduce redundancy. Hence, we have to develop a technique that selects a band subset with high relevance and high significance. In this paper, the Rough set theory is used to select the relevant bands from high dimensional hyperspectral data.

Let $U = \{x_1, x_2, ..., x_n\}$ be the set of n available labeled patterns and $B = \{b_1, b_2, ..., b_m\}$ be the set of m bands of hyperspectral data. We can represent these patterns as $W = \{w_{ij} \mid i = 1, ..., n; j = 1, ..., m\}$, where w_{ij} is the value of the pattern $x_i \in U$ associated with band $b_j \in B$. Let D be a decision attribute that contains the class label of the pattern $x_i \in U$. Accordingly, in terms of Rough set theory, a hyperspectral data set can be represented by a decision table $I = (U, B \cup D)$, where B and D play the role of condition and decision attribute sets, respectively.

In hyperspectral remote sensing data, the class labels of pixels/regions are represented by discrete values, while the band/feature values of the pixels/patterns are continuous. In order to exploit Rough set theory for band selection, the continuous band values of the patterns should be divided into several discrete values to generate equivalence classes [40], [41]. Different discretization methods can be employed to discretize the continuous band values, such as equal width interval binning or equal frequency binning [41], mean and standard deviation based discretization method [38], [41], roughfication method [42], etc. In this work we employed the simple equal width interval binning approach (uniform quantization) [41]. It does not make use of any class membership information during the discretization process. The uniform quantization algorithm determines the minimum and maximum values of the discretized attribute and then divides the range into a user-defined number of equal width discrete intervals.

After discretization of continuous attributes (spectral bands), the relevance and significance of the bands are computed using the Rough set theory. In terms of Rough set theory, the relevance $r(b_i, D)$ of a hyperspectral band b_i with respect to the decision attribute D can be computed using (4), i.e.

$$r(b_i, D) = \gamma_{b_i}(D) \tag{6}$$

If we select $k (k < m)$ spectral bands based only on the relevance criterion defined in (6), we may have high redundancy. Let us assume that b_i and b_j are two highly correlated bands selected due to their high relevance values. Since the two selected bands are highly correlated, the classification prediction accuracy would not change significantly if one of them is removed. It follows that one band is not useful with respect to the other. The significance criterion defined in (5) is able to find out the irrelevant bands. If the significance of a band with respect to another band is 0, then the band is completely irrelevant. Thus, the significance criterion play an important role to select informative bands. In terms of Rough set theory, the significance $z(b_i, b_j)$ of a band b_j with respect

to another band b_i can be computed using (5), i.e.

$$z(b_i, b_j) = \delta_{\{b_i,b_j\}}(D, b_j) = \gamma_{\{b_i,b_j\}}(D) - \gamma_{b_j}(D). \quad (7)$$

In our proposed method, we selects a subset S of k most informative bands from the set B of m bands by taking into an account the relevance and significance criteria. When $k = 1$, the method computes the relevance of each band $b_i \in B$ using (6). Then the band that has the highest relevance value is selected as the most informative band. When $k > 1$, the first-order incremental search is used to select one band at each time [38], [43]. In first-order search, it is assumed that the $(k-1)$ bands are already selected. The k^{th} informative band from the set $B - S$ (the difference between sets B and S is denoted as $B - S$) is chosen based on a function which incorporates relevance and significance criteria. In the literature this function is defined as follows [38], [39]:

$$F(b_j) = r(b_j, D) + \frac{1}{\mid S \mid} \sum_{b_i \in S} z(b_i, b_j), \quad (8)$$

where $b_j \in (B-S)$. b_l is selected as the k^{th} informative band if it produces the largest value of the function defined in (8), i.e., $b_l = \arg \max_{b_i \in (B-S)} \{F(b_i)\}$. The function in (8) has two terms, which are associated with relevance and significance criteria, respectively. If a band in $(B-S)$ has zero significance value with one band in S and very high significance value with other bands in S, then the second term contributes a large value in (8). As a result, the band may be selected as an informative band, although it is completely redundant. To mitigate this limitation, in this article we modify the second term of (8) by defining the following function:

$$F^{'}(b_j) = r(b_j, D) + \frac{\min\{z(b_i, b_j)\}}{\max\{z(b_i, b_j)\}} \min\{z(b_i, b_j)\}, \quad (9)$$

where $b_i \in S$ and $b_j \in (B - S)$. Here one can see that the significance term corresponding to a band completely depends on its minimum significance value, which seems more reasonable to identify an informative band. Unlike the techniques presented in [27]–[34], [37], the proposed technique avoids the generation of a large number of reducts by adopting a first-order incremental searching algorithm. This makes it less computationally demanding. Note that, like the minimum-redundancy-maximum-relevance (mRMR) based feature selection method [43], the proposed method selects subset of features from the entire feature set by maximizing the relevance and minimizing the redundancy of the selected features. However, the proposed technique exploits Rough sets theory instead of mutual information as used by the mRMR technique for relevance and redundancy measures. The major advantage of the proposed method is that the redundancy measure (significance criterion) of the mRMR method does not take into account the class labels information, while both relevance and redundancy measures of the proposed method are computed based on the class label information. The complete algorithm of the proposed technique is given below:

Algorithm 1 Proposed band selection method

1: Initialize $S = \phi$ and $B = \{b_1, b_2, ..., b_m\}$.
2: Compute the relevance of each band $b_i \in B$ using (6).
3: Select the band $b_j \in B$ that has the maximum relevance value, i.e., $b_j = \arg \max_{b_i \in B} \{r(b_i, D)\}$.
4: Repeat the following three steps until the desired number of bands are selected.
5: Update $S = \{S \cup b_j\}$ and $B = \{B - b_j\}$.
6: Compute the significance of each of the remaining bands in B with respect to the selected bands in S using (7).
7: Select $b_j \in B$ that results in the maximum value in (9).

The band selection technique presented in algorithm (1) has low computationally complexity. The time complexity to compute the relevance values of m bands with n labeled patterns is $O(mn)$. Then the complexity of selecting the most relevant band among m features is $O(m)$. The time complexity to compute the significance value of a band with respect to the $(d-1)$ already selected features is $O(d-1)$. So, the time taken to compute significance values for selecting d bands will be $O((d-1)(m-d+1)) = O(dm)$. Accordingly, the time complexity of algorithm (1) is $O(nm)+O(m)+O(dm) = O(nm)$.

IV. Description of data sets

In order to assess the effectiveness of the proposed band selection method, experiments were carried out on three hyperspectral data sets of the Kennedy Space Center (KSC), Florida [44], the Okavango Delta, Botswana [44] and the Indian Pine test site of Northwestern Indiana [45].

The first data set is a hyperspectral image acquired by AVIRIS sensor on the Kennedy Space Center (KSC), Merritt Island, Florida, USA, on March 23, 1996 (see Fig. 1). This image consists of 512 x 614 pixels and 224 bands with a spatial resolution of 18 m. The number of bands is initially reduced to 176 by removing water absorption and low signal-to-noise channels. The available labeled data were collected using land-cover maps derived from color infrared photography provided by KSC and Landsat Thematic Mapper imagery. The reader is referred to [44] (or to http://www.csr.utexas.edu/hyperspectral) for more details on this data set. After the elimination of noisy samples, the available labelled samples are used for the experiment. The class names and corresponding numbers of ground truth observations used in the experiments are listed in Table II.

The second data set is a hyperspectral image acquired by the Hyperion sensor on the Okavango Delta, Botswana, on May 31, 2001 (see Fig. 2). The Hyperion sensor on EO-1 acquires data at 30 m/pixel resolution over a 7.7km strip in 242 bands covering the 400-2500 nm portion of the spectrum in 10 nm windows. The pre-processing of the data was performed by the University of Texas Center for Space Research. Uncalibrated and noisy bands were removed, and the remaining 145 bands were included as candidate features ([10-55], [82-97], [102-119], [134-164], [187-220]). The fourteen identified classes represent the land-cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of

Fig. 1. False colour composition of the hyperspectral KSC image.

TABLE II
KSC DATA SET: CLASS NAMES AND NUMBER OF SAMPLES.

Class Name	No. of Samples
Scrub	761
Willow swamp	241
Cabbage palm hammock	256
Cabbage palm/Oak hammock	251
Slash pine	161
Oak/Broadleaaf hammock	229
Hardwood swamp	105
Graminoid marsh	431
Spartina marsh	520
Cattaial marsh	377
Salt marsh	419
Mud flats	462
Water	908

the delta. These classes were chosen to reflect the impact of flooding on vegetation in the study area. The class names and corresponding numbers of ground truth observations used in the experiments are listed in Table III. The reader is referred to [44] (or to http://www.csr.utexas.edu/hyperspectral) for more details on this data set.

TABLE III
BOTSWANA DATA SET: CLASS NAMES AND NUMBER OF SAMPLES.

Class Name	No of Samples
Water	270
Hippo grass	101
FloodPlain grasses 1	251
FloodPlain grasses 2	215
SReeds	269
Riparian	269
Firescar	259
Island interior	203
Acacia woodlands	314
Acacia shrublands	248
Acacia grasslands	305
Short mopane	181
Mixed mopane	268
Exposes soils	95

The third data set is another hyperspectral image acquired by the AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) sensor over the agricultural land of Indian Pine, Indiana in the early growing season of 1992 (see Fig. 3). These data

Fig. 2. False colour composition of the hyperspectral Botswana image.

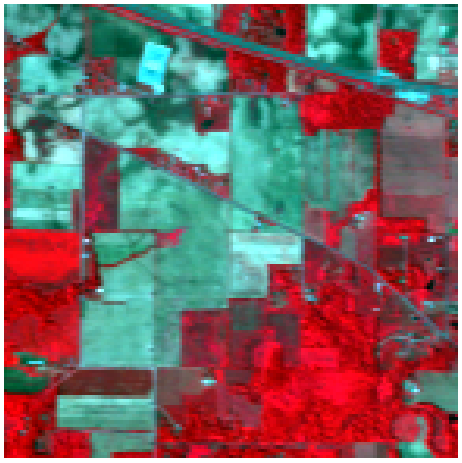

Fig. 3. False colour composition of the hyperspectral Indian Pine image.

TABLE IV
INDIAN PINE DATA SET: CLASS NAMES AND NUMBER OF SAMPLES.

Class Name	No of Samples
Alfalfa	46
Corn-notill	1428
Corn-min	830
Corn	237
Grass/Pasture	483
Grass/Trees	730
Grass/Pasture-mowed	28
Way-windrowed	478
Oats	20
Soybeans-notill	972
Soybeans-min	2455
Soybean-clean	593
Wheat	205
Woods	1265
Bldg-Grass-Tree-Drives	386
Stone-steel towers	93

were acquired in the spectral range 400-2500 nm with spectral resolution of about 10 nm. The image consists of 145 x 145 pixels and 220 spectral bands with a spatial resolution of 20 m. Twenty water absorption and fifteen noisy bands were removed and the remaining 185 bands were included as candidate features ([4-102], [113-147], [166-216]). The class names and corresponding numbers of ground truth observations used in the experiments are listed in Table IV. The reader is referred to [45] (or to http://engineering.purdue.edu/ biehl/MultiSpec) for more details on this data set.

V. Experimental results

A. Design of experiments

To assess the effectiveness of the proposed band selection method we compared it with three other state-of-the-art methods: i) the supervised Rough set and fuzzy C-Means (referred as Rough-FCM) [11]; ii) the unsupervised wards linkage strategy using divergence (referred as WaLuDi) [16]; and iii) the unsupervised wards linkage strategy using mutual information (referred as WaLuMI) [16]. The Rough-FCM approach, first computes a relevance vector corresponding to each band of the hyperspectral image. The components of the k^{th} relevance vector indicate the relevance of the different class objects represented by the k^{th} band. Thus, a hyperspectral image with m bands and c objects will generate m relevance vectors with c components each. After generation of the relevance vectors, fuzzy C-Means clustering is applied to them to group similar bands into a cluster. Finally, one band from each cluster with maximum grade of fuzzy membership is selected [11]. In the experiments, for all the considered data sets, we selected randomly 50% of the available labeled samples for defining the training set given as input to both the proposed and the Rough-FCM supervised band selection techniques. The WaLuDi and WaLuMI are unsupervised approach. Both adopts a clustering technique to group similar bands into a cluster. Then from each cluster, the WaLuDi and the WaLuMI selected the most informative bands by applying the Kullback-Leibler divergence and mutual information criterion, respectively [16].

Before conducting the experiments, normalization of spectral channels (scaling them between 0 and 1) was performed. For both the data sets, the discretization of the proposed technique is done by fixing the number of bins to 100. The desired number of bands to be selected is not known a priori and varied in the different images. In the present investigation, experiments were carried out for different numbers of bands ranging from 5 to 40 with a step size of 5. After the termination of the band selection algorithm, to evaluate the effectiveness of selected bands, we adopted an one-against-all (OAA) architecture of support vector machine (SVM) classifiers. Each SVM was implemented with radial basis function (RBF) kernels. The SVM parameters $\{\sigma, C\}$ (the spread of the RBF kernel and the regularization parameter) for all the data sets were derived by applying a grid search according to a ten-fold cross-validation technique. The cross-validation procedure aimed at selecting the parameter values for the SVM. For all the data sets, 50% of the available labeled samples (shown in Sec. IV) were randomly selected and included in the training set used for the learning of the SVM classifier. Then the accuracy was evaluated on the remaining test samples. To reduce the random effect of the results, 10 trials with different training sets were performed and the average results were reported.

The multiclass SVM with the OAA architecture has been implemented by using the LIBSVM library (for Matlab interface) [46]. All the methods presented in this paper have been implemented in Matlab.

B. Results: KSC data set

The first experiment was carried out to compare the performance of the proposed technique with other state-of-the-art techniques by using the KSC hyperspectral data set described in Section IV. Fig. 4 shows the average overall classification accuracies provided by different band selection methods versus the number of bands. From this figure, one can see that the proposed technique produced superior results as compared to the considered literature techniques especially when small number of bands are selected. The WaLuDi, the WaLuMI and the Rough-FCM techniques grouped the similar bands into a cluster to remove the redundant channels. When the number of clusters is very small, these techniques may fail to distribute the informative bands into different clusters. On the other hand, the proposed technique searches the informative bands from the whole band/feature space. As a result, for a small number of selected bands the literature techniques provided poor results compared with the proposed technique. For a quantitative analysis, Table V reports the average overall classification accuracies, as well as the average kappa accuracies obtained on 10 runs for the hyperspectral KSC data set. From the table, one can see that by selecting only 5 bands, the proposed technique was able to achieve 85.03% overall accuracy. The WaLuDi, the WaLuMI and the Rough-FCM techniques with the same number of selected bands resulted in an overall accuracy of 82.63%, 80.67% and 81.18%, respectively. It can also be seen that selecting only 20 bands, the proposed technique was able to achieve classification accuracy similar to that produced by the WaLuDi method by selecting 40 bands. This confirms the effectiveness of the proposed technique for the KSC data set. Table VI shows the first 10 bands selected by different approaches.

TABLE V
Average (obtained on ten runs) overall classification accuracy ($\bar{OA}$) and kappa accuracy (KSC data set)

No of Bands	WaLuDi		WaLuMI		Rough-FCM		Proposed	
	$\bar{OA}$	kappa	$\bar{OA}$	kappa	$\bar{OA}$	kappa	$\bar{OA}$	kappa
5	82.63	.806	80.67	.783	81.18	.790	**85.03**	**.833**
10	85.71	.841	88.83	.876	87.52	.861	**90.51**	**.894**
15	89.04	.878	92.25	.914	92.22	.913	**93.01**	**.922**
20	89.65	.885	93.15	.924	92.32	.914	**93.73**	**.930**
25	90.64	.896	93.66	.929	93.63	.929	**94.39**	**.937**
30	91.62	.907	94.23	.936	94.41	.938	**94.66**	**.941**
35	92.17	.913	94.52	.939	**94.97**	**.944**	94.93	.944
40	93.70	.930	95.11	.946	95.11	.946	**95.23**	**.947**

C. Results: Botswana data set

The second experiment was carried out to compare the performance of the proposed technique with the other considered state-of-the-art techniques by using the Botswana data

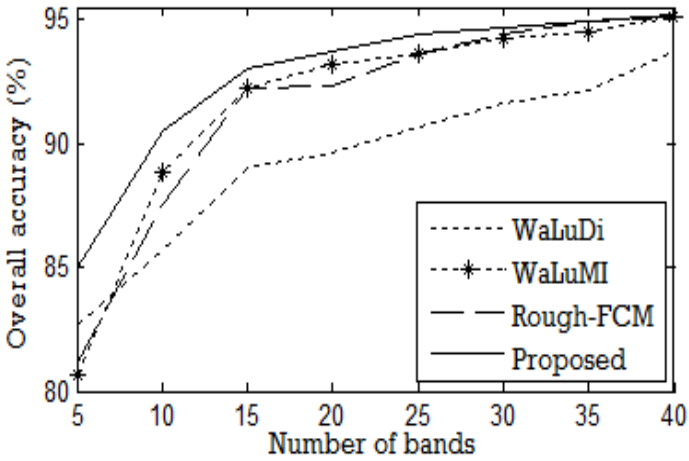

Fig. 4. Average overall classification accuracy (over ten runs) versus the number of selected bands provided by the WaLuDi, the WaLuMI, the Rough-FCM, and the Proposed methods (KSC data set).

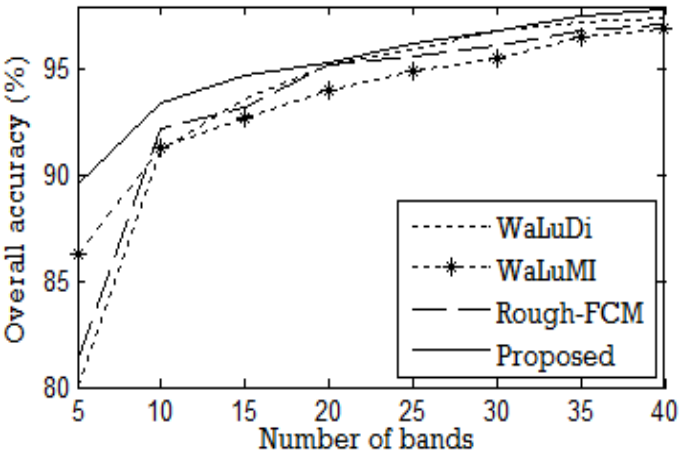

Fig. 5. Average overall classification accuracy (over ten runs) versus the number of selected bands provided by the WaLuDi, the WaLuMI, the Rough-FCM, and the Proposed methods (Botswana data set).

TABLE VI
THE TEN MOST INFORMATIVE BANDS SELECTED BY THE WALUDI, THE WALUMI, THE ROUGH-FCM AND THE PROPOSED METHOD (KSC DATA SET)

Methods	Selected bands
WaLuDi	7, 21, 43, 48, 89, 173, 187, 200, 201, 202
WaLuMI	19, 27, 45, 67, 75, 89, 134, 140, 187, 195
Rough-FCM	27, 29, 35, 73, 83, 120, 131, 175, 195, 199
Proposed	34, 36, 37, 38, 40, 49, 59, 76, 88, 134

TABLE VII
AVERAGE (OBTAINED ON TEN RUNS) OVERALL CLASSIFICATION ACCURACY ($\bar{OA}$), AND KAPPA ACCURACY (BOTSWANA DATA SET)

No of Bands	WaLuDi		WaLuMI		Rough-FCM		Proposed	
	$\bar{OA}$	kappa	$\bar{OA}$	kappa	$\bar{OA}$	kappa	$\bar{OA}$	kappa
5	80.01	.783	86.30	.852	81.32	.798	**89.63**	**.888**
10	91.20	.905	91.35	.906	92.19	.915	**93.44**	**.929**
15	93.66	.931	92.72	.921	93.22	.927	**94.70**	**.943**
20	95.19	.948	94.05	.936	95.24	.948	**95.30**	**.949**
25	95.96	.956	94.94	.945	95.68	.953	**96.22**	**.959**
30	96.80	.965	95.55	.952	96.18	.959	**96.87**	**.966**
35	97.20	.970	96.54	.962	96.88	.966	**97.57**	**.974**
40	97.46	.973	96.93	.967	97.13	.969	**97.80**	**.976**

set described in Section IV. Fig. 5 shows the average overall classification accuracies provided by different feature selection methods versus the number of bands. From this figure, one can see that for small numbers of bands, the proposed technique produced significantly higher classification accuracy as compared to the WaLuDi, the WaLuMI and the Rough-FCM techniques. When the number of selected bands increased it provided slightly better accuracies than the existing techniques. For a quantitative analysis, Table VII reports the average overall classification accuracies, as well as the average kappa accuracies, obtained on 10 runs. From the table, one can see that by selecting only 5 bands, the proposed technique was able to produce 89.63% overall accuracy. The WaLuDi, the WaLuMI and the Rough-FCM techniques resulted in an overall accuracy of 80.01%, 86.30%, and 81.32%, respectively. It can also be seen that the proposed technique resulted in sharply higher accuracies than the supervised Rough-FCM method. This is because the proposed technique exploits class label information properly in order to select the informative bands. The results confirm the effectiveness of the proposed technique also on the Botswana data set. Table VIII shows the first 10 bands selected by different approaches.

D. Results: Indian Pine data set

The third experiment was carried out to compare the performance of the proposed technique with those of the other considered state-of-the-art techniques by using the Indian Pine data set described in Section IV. Fig. 6 shows the average overall classification accuracies provided by different methods. From this figure, one can see that for small numbers of bands, the proposed technique produced higher classification accuracy as compared to other existing techniques. For higher number of selected bands, the proposed technique provided similar accuracies as compared to the results produced by the best state-of-the-art method. For a quantitative analysis, Table IX reports the average overall classification accuracies, as well as the average kappa accuracies, obtained on 10 runs. From the table, one can see that by selecting only 5 bands, the proposed technique resulted in an overall accuracy of 67.19% overall accuracy. The WaLuDi, the WaLuMI and the Rough-FCM techniques resulted in an overall accuracy of 60.07%, 60.09%, and 63.92%, respectively. Table X shows the first 10 bands selected by the different approaches.

The last experiment was devoted to analyze the performance of the proposed technique by varying the width of the bins used to discretize the continuous band values of the patterns. To this end, for the KSC, the Botswana and the Indian Pine data sets, the number of bins was varied in the range 60, 70, 80, 90, and 100. Table XI shows the average classification accuracies obtained by 10 bands, selected with different bin widths. By analyzing these results, one can conclude that the accuracy of the proposed method does not significantly change in a wide range of bin widths.

TABLE VIII
THE TEN MOST INFORMATIVE BANDS SELECTED BY THE WALUDI, THE WALUMI, THE ROUGH-FCM AND THE PROPOSED METHOD (BOTSWANA DATA SET)

Methods	Selected bands
WaLuDi	11, 18, 25, 35, 50, 91, 96, 97, 150, 164
WaLuMI	11, 28, 44, 88, 110, 140, 158, 193, 207, 212
Rough-FCM	18, 31, 41, 91, 147, 160, 190, 209, 211, 212
Proposed	35, 36, 38, 45, 92, 106, 119, 135, 150, 196

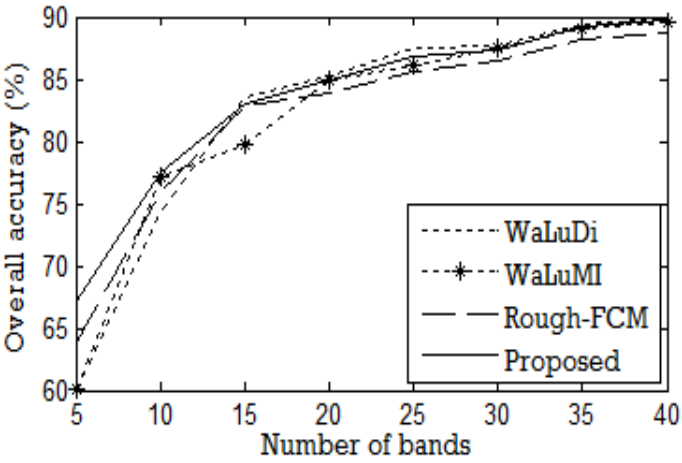

Fig. 6. Average overall classification accuracy (over ten runs) versus the number of selected bands provided by the WaLuDi, the WaLuMI, the Rough-FCM, and the Proposed methods (Indian Pine data set).

TABLE IX
AVERAGE (OBTAINED ON TEN RUNS) OVERALL CLASSIFICATION ACCURACY ($\bar{OA}$), AND KAPPA ACCURACY (INDIAN PINE DATA SET)

No of Bands	WaLuDi		WaLuMI		Rough-FCM		Proposed	
	$\bar{OA}$	kappa	$\bar{OA}$	kappa	$\bar{OA}$	kappa	$\bar{OA}$	kappa
5	60.07	.529	60.09	.531	63.92	.579	**67.19**	**.618**
10	74.35	.704	77.16	.737	76.01	.724	**77.53**	**.742**
15	**83.47**	**.811**	79.67	.766	82.77	.803	83.02	.806
20	**85.31**	**.832**	84.99	.828	83.94	.816	84.85	.826
25	**87.42**	**.856**	86.10	.841	85.65	.836	86.73	.848
30	**87.74**	**.860**	87.42	.856	86.41	.845	87.41	.856
35	**89.47**	**.880**	89.12	.876	88.15	.865	89.14	.876
40	**89.83**	**.884**	89.49	.880	88.71	.871	89.66	.882

TABLE X
THE TEN MOST INFORMATIVE BANDS SELECTED BY THE WALUDI, THE WALUMI, THE ROUGH-FCM AND THE PROPOSED METHOD (INDIAN PINE DATA SET)

Methods	Selected bands
WaLuDi	31, 52, 56, 58, 69, 84, 129, 142, 171, 197
WaLuMI	24, 42, 49, 72, 97, 122, 142, 182, 191, 209
Rough-FCM	7, 17, 22, 30, 59, 79, 118, 125, 167, 182
Proposed	29, 35, 44, 77, 84, 130, 166, 182, 214, 215

TABLE XI
AVERAGE CLASSIFICATION ACCURACY PROVIDED BY TEN BANDS SELECTED WITH DIFFERENT BIN WIDTH

No of bins	Data sets		
	KSC	Botswana	Indian Pine
60	90.69	93.21	78.69
70	90.05	93.07	77.97
80	90.35	93.38	77.83
90	90.17	93.59	78.15
100	90.51	93.44	77.53

VI. DISCUSSION AND CONCLUSION

In this paper we presented a supervised method based on the Rough set theory for hyperspectral band selection. Our technique selects a subset of informative bands from hyperspectral imagery by using relevance and significance criteria. The Rough set theory is exploited to compute the relevance and significance values of each of the hyperspectral band. Moreover, a novel criterion is proposed to select the most informative bands. The proposed technique uses simple first-order incremental search to avoid the generation of large numbers of reducts, thus resulting less computational demands as compared to the conventional Rough set based approaches.

To assess the effectiveness of the proposed band selection technique, we compared it with one Rough set based supervised technique and two unsupervised techniques based on KL divergence and mutual information criterion existing in the remote sensing literature by using three hyperspectral data sets. The results of this comparison pointed out that for small numbers of selected bands, the proposed method always provided significantly higher accuracies compared to all the reference band selection methods. For larger numbers of selected bands, it produced similar accuracy as compared to the best result obtained by the existing state-of-the art methods.

As future developments of this work, we plan to define a strategy for selecting a band subset by detecting the class-wise most informative bands to further improve results.

REFERENCES

[1] C.-I. Chang, Q. Du, T.-L. Sun, and M. L. G. Althouse, "A joint band prioritization and band decorrelation approach to band selection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2631–2641, 1999.

[2] X. R. Wang, A. J. Brown, and B. Upcroft, "Applying incremental EM to Bayesian classifiers in the learning of hyperspectral remote sensing data," in *Proceedings of 7th International Conference on Information Fusion*, 2005, pp. 606–613.

[3] P. K. Varshney and M. K. Arora, *Advanced Image Processing Techniques for Remotely Sensed Hyperspectral Data*. Berlin: Springer-Verlag, 2004.

[4] T. A. Warner and M. C. Shank, "Spatial autocorrelation analysis of hyperspectral imagery for feature selection," *Remote Sens. Environment*, vol. 60, pp. 58–70, 1997.

[5] L. Bruzzone, F. Roli, and S. B. Serpico, "An extension of the jeffreys-matusita distance to multiclass cases for feature selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 6, pp. 1318–1321, 1995.

[6] L. Bruzzone and S. B. Serpico, "A technique for feature selection in multiclass problems," *Int. J. Remote Sens.*, vol. 21, no. 3, pp. 549–563, 2000.

[7] S. B. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1360–1367, 2001.

[8] A. Ifarraguerri and M. W. Prairie, "Visual method for spectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 2, pp. 101–106, 2004.

[9] R. Huang and M. He, "Band selection based on feature weighting for classification of hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 2, pp. 156–159, 2005.

[10] S. D. Backer, P. Kempeneers, W. Debruyn, and P. Scheunders, "A band selection technique for spectral classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 3, pp. 319–323, 2005.

[11] H. Shi, Y. Shen, and Z. Liu, "Hyperspectral bands reduction based on rough sets and fuzzy c-means clustering," in *Instrumentation and Measurement Technology Conference (IMTC 2003), USA*, vol. 2, 2003, pp. 1053–1056.

[12] L. Bruzzone and C. Persello, "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 9, pp. 3180–3191, 2009.

[13] H. Yang, Q. Du, H. Su, and Y. Sheng, "An efficient method for supervised hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 138–142, 2011.

[14] C.-I. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1575–1585, 2006.

[15] Q. Du and Y. H., "Similarity-based unsupervised band selection for hyperspectral image analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 564–568, 2008.

[16] A. Martinez-Uso, F. Pla, J. M. Sotoca, and P. Garcia-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4158–4171, 2007.

[17] C. Yang, S. Liu, L. Bruzzone, R. Guan, and P. Du, "A feature-metric-based affinity propagation technique for feature selection in hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 5, pp. 1152–1156, 2013.

[18] B. Guo, R. I. Damper, S. R. Gunn, and J. D. B. Nelson, "A fast separability-based feature-selection method for high-dimensional remotely sensed image classificaion," *Pattern Recog.*, vol. 41, no. 5, pp. 1653–1662, 2008.

[19] A. Das, S. Ghosh, and A. Ghosh, "Band elimination of hyperspectral imagery using partitioned band image correlation and capacitory discrimination," *Int. J. Remote Sens.*, vol. 35, no. 2, pp. 554–577, 2014.

[20] B. C. Kuo and D. A. Landgrebe, "Nonparametric weighted feature extraction for classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 5, pp. 1096–1105, 2004.

[21] A. J. Brown, "Spectral curve fitting for automatic hyperspectral data analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1601–1608, 2006.

[22] S. B. Serpico and G. Moser, "Extraction of spectral channels from hyperspectral images for classification purposes," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 2, pp. 484–494, 2007.

[23] B. Mojaradi, H. A. Moghaddam, M. J. V. Zoej, and R. P. W. Duin, "Dimensionality reduction of hyperspectral data via spectral feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2091–2105, 2009.

[24] J. M. Yang, P. T. Yu, and B. C. Kuo, "A nonparametric feature extraction and its application to nearest neighbour classification for hyperspectral image data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1279–1293, 2010.

[25] R. R. Marpu, M. Pedergnana, M. Dalla Mura, S. Peeters, J. A. Benediktsson, and L. Bruzzone, "Classification of hyperspectral data using extended attribute profiles based on supervised feature extraction," *Int. J. Image and Data Fusion*, vol. 3, no. 3, pp. 269–298, 2012.

[26] M. Imani and H. Ghassemian, "Band clustering-based feature extraction for classification of hyperspectral images using limited training samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 8, pp. 1325–1329, 2014.

[27] Z. Pawlak, *Rough Sets, Theoretical Aspects of Resoning About Data*. Kluwer, Dordrecht, The Netherlands, 1991.

[28] J. Komorowski, Z. Pawlak, L. Polkowski, and A. Skowron, "Rough sets: a tutorial," in *Rough-Fuzzy Hybridization: A New Trend in Decision Making*, S. K. Pal and A. Skowron, Eds. Springer Verlag, Singapore, 1999, pp. 3–98.

[29] A. A. Skowron, R. W. Swiniarski, and P. Synak, "Approximation spaces and information granulation," *Trans. Rough Set*, vol. 3, pp. 175–189, 2005.

[30] R. Jensen and Q. Shen, "Fuzzy-rough attribute reduction with application to web categorization," *Fuzzy Sets and Systems*, vol. 141, no. 3, pp. 469–485, 2004.

[31] ——, "Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approach," *IEEE Trans. Knowledge and Data Engineering*, vol. 16, no. 12, pp. 1457–1471, 2004.

[32] D. Slezak, "Rough sets and few-objects-many-attributes problem: the case study of analysis of gene expression data sets," in *Proceedings of the Frontiers in the Convergence of Bioscience and Informaion Technologies*, 2007, pp. 233–240.

[33] A. A. Chouchoulas and Q. Shen, "Rough set-aided keyword reduction for text categorisation," *Applied Artificial Intelligence*, vol. 15, pp. 843–873, 2001.

[34] C. Cornelis, R. Jensen, G. H. Martin, and D. Slezak, "Attribute selection with fuzzy decision reducts," *Information Sciences*, vol. 180, pp. 209–224, 2010.

[35] A. Skowron and C. Rauszer, "The discernibility matrices and functions in information systems," in *Intelligent Decision Support*, R. Slowinski, Ed. Kluwer Academic Publishers, Dordrecht, 1992, pp. 331–362.

[36] J. Wroblewski, "Finding minimal reducts using genetic algorithms," in *Proceedings of the 2nd Annual Joint Conference on Information Sciences*, 1995, pp. 186–189.

[37] N. N. Zhong, J. Dong, and S. Ohsuga, "Using rough sets with heuristics for feature selection," *Journal of Intelligence System*, vol. 16, pp. 199–214, 2001.

[38] P. Maji and S. Paul, "Rough sets for selection of molecular descriptors to predict biological activity of molecules," *IEEE Transactions on Systems, Man, and CyberneticsPart C: Applications and Reviews*, vol. 40, no. 6, pp. 639–648, 2010.

[39] ——, "Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data," *International Journal of Approximate Reasoning*, vol. 52, pp. 408–426, 2013.

[40] M. J. Beynon, "Stability of continuous value discretisation: an application within rough set theory," *International Journal of Approximate Reasoning*, vol. 35, pp. 29–53, 2004.

[41] S. Kotsiantis and D. Kanellopoulos, "Discretization techniques: A recent survey," *GESTS International Transactions on Computer Science and Engineering*, vol. 32, no. 1, pp. 47–58, 2006.

[42] D. Slezak and J. Wroblewski, "Roughfication of numeric decision tables: the case study of gene expression data." Springer, Berlin, Heidelberg, 2007, pp. 316–323.

[43] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of Max-Dependence, Max-Rrelevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[44] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, 2005.

[45] L. O. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2653–2667, 1999.

[46] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machine*, 2001, software available at http://csie.ntu.edu.tw/čjlin/libsvm.