Subspace Detection Using a Mutual Information Measure for Hyperspectral Image Classification

Md. Ali Hossain, Xiuping Jia, *Senior Member, IEEE*, and Mark Pickering, *Member, IEEE*

***Abstract*—Finding a subspace which consists of the most informative features for reliable hyperspectral image classification is a challenging task. Feature reduction is often achieved via feature selection and feature extraction techniques. In this letter, a hybrid approach which combines both treatments is proposed. Principal Component Analysis (PCA) is applied as a preprocessing step so that each of the new features is generated from the complete set of the original spectral bands. Feature selection is then performed effectively using a normalized Mutual Information (nMI) measure with two constraints to maximize general relevance and minimize redundancy in the selected subspace. The proposed algorithm (PCA-nMI) is tested on hyperspectral images and the experimental results show that the modifications give significant improvement in terms of classification accuracy.**

***Index Terms*—Feature extraction, feature selection, hyperspectral image, mutual information, principal components.**

I. Introduction

With the advent of hyperspectral remote sensors, hundreds of narrow contiguous spectral bands can now be captured to provide greater details on the spectral variation of targets than with conventional multispectral systems. However, at present, a major challenge on the use of hyperspectral images lies in the development of reliable and effective techniques for processing the large amounts of data. For instance, the AVIRIS sensor simultaneously measures 224 bands with a fine spectral resolution of 0.01 μm. Such a large number of bands implies high-dimensionality data in a machine learning framework. If the training samples are limited, a reduction in the classification accuracy of the test data is often observed due to the poor generalization of the training results and this effect is known as the Hughes phenomenon [1]. On the other hand, since the hyperspectral data are sensed in very close and contiguous spectral bands, some bands are highly correlated and can be removed. It can also be observed that not all the bands are important for a specific application. Therefore, the aforementioned problem can be solved by feature reduction via feature selection or feature extraction [2], [3]. Generating a new set of K features from the original N features/bands $(K < N)$ is referred to as feature extraction, which could be obtained via linear or nonlinear transforms with supervised or unsupervised approaches. For example Linear Discriminant Analysis (LDA) [4] and Principal Component Analysis (PCA) [5] are well-known supervised and unsupervised linear feature extraction methods, respectively. The LDA uses the first and second order statistics of the data to find the largest ratio of between class means and within class variance in the new subspace. A limitation of the LDA approach is that it can only generate $M-1$ new features, where M is the number of classes of interest. It is also limited to handle the case where each class forms a single, normally distributed cluster. The PCA algorithm can represent the data optimally in terms of maximal variance and minimal mean square error between the new representation given by a reduced number of components and the original data. The new features are also not correlated. However, PCA does not emphasize individual classes during eigen analysis, so there is no guarantee that the selected features are the best for class separation [5]–[7]. Since PCA considers global statistics, there may be some classes which are small and hardly affect the overall statistics and hence are not represented in the first few components. A popular supervised feature selection approach is based on the Jeffries-Matusita (J-M) distance which measures the separability between a pair of classes after each of them is assumed to have a Gaussian distribution [8]. The limitations and drawbacks of this method, however, include a strong dependence on the training data, ineffective class pair wise treatment and the assumption of normally distributed class data [8], [9]. Alternatively, Mutual Information (MI) has been used to select relevant features [10] and for band clustering [11], [12] by measuring both linear and nonlinear relationships between the spectral bands and the target classes. This approach is nonparametric, works for non-Gaussian data and is able to handle multiclass cases without the need for individual class pair evaluations. Multiple feature selection techniques using MI have been developed, such as MIFS [13] and mRMR [14]. In this letter, a weakness of the current MI based methods is identified, and modifications are proposed in order to guarantee only informative features are selected. Our proposed subspace detection method combines the use of PCA and normalized MI (PCA-nMI).

Manuscript received June 21, 2012; revised December 17, 2012 and March 5, 2013; accepted April 5, 2013. Date of publication August 15, 2013; date of current version November 25, 2013.

The authors are with the School of Engineering and Information Technology, University of New South Wales, Canberra BC 2610, Australia (e-mail: Md.Hossain3@student.adfa.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2013.2264471

II. Proposed Feature Reduction Method

A. Feature Selection Based on Mutual Information

The mutual information (MI) between two input images $\mathbf{X}$ and $\mathbf{Y}$ measures general dependency between them and is defined as

$$I(\mathbf{X},\mathbf{Y}) = \sum_{y=1}^{L_1}\sum_{x=1}^{L_2} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

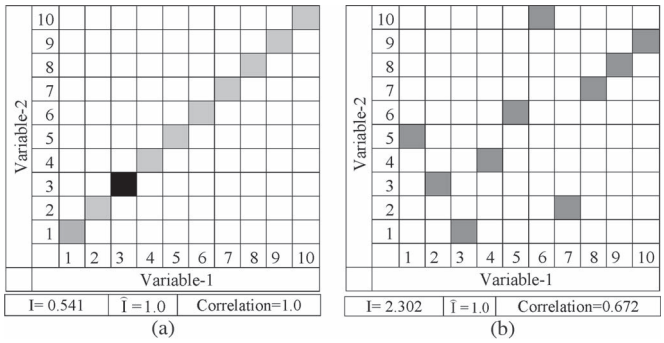

Fig. 1. Scatter plots of (a) perfect linear relationship and (b) point-to-point nonlinear relationship between two variables. Darkness represents the value of frequency.

where x, y are the pixel values and L_1, L_2 are the maximum pixel values of the input images. From the definition, MI is bounded by $H(\mathbf{X})$, the entropy of $\mathbf{X}$, if $H(\mathbf{X}) < H(\mathbf{Y})$. The application of MI has several benefits that distinguish it from other feature selection criteria. The variables can be indices or categories, such as class labels. It has the ability to measure general relationships including both linear and nonlinear [15]. Fig. 1 gives the scatter plots of two example data sets where the darkness represents the value of frequency. The data set shown in Fig. 1(a) has a perfect linear relationship, and the data set shown in Fig. 1(b) has a perfect one to one, but nonlinear relationship (only one entry in each row and each column). The MI values of both cases are the highest, which are the entropy values of the two data sets, respectively, while Pearson's correlation coefficient is 1 and 0.672, respectively. It can be seen that correlation is high when two variables have a perfect linear relationship but low when the relationship is nonlinear. In general, when MI is low (a normalized MI will be introduced in the next section to make this judgement easier), correlation will be low, but if MI is high, correlation may not be high (nonlinear case). The MI can be used as a selection criterion during supervised feature selection to identify the most relevant features when the input is training data in the i^{th} band, $\mathbf{X}_i$, and $\mathbf{Y}$ is the class labels $\mathbf{C}$ of the training data. The features having a higher MI value with the classification objective are more suitable to use for classification [16]. Two bounds (lower and upper) on Bayes error justify the benefits of using MI. Using Fano's inequality [15] we can find that the probability of incorrect estimation P_e of class data $\mathbf{C}$ for the given i^{th} input feature $\mathbf{X}_i$ is lower bounded by

$$P_e \geq \left(H(\mathbf{C}) - I(\mathbf{X}_i, \mathbf{C})\right) / \log N_c \tag{2}$$

and upper bounded by

$$P_e \leq \frac{1}{2}\left(H(\mathbf{C}) - I(\mathbf{X}_i, \mathbf{C})\right) \tag{3}$$

The entropy $H(\mathbf{C})$ is calculated from the histogram of the numeric values of the class labels for the training data. Since the number of classes N_c and the entropy $H(\mathbf{C})$ are fixed, P_e is minimized when $I(\mathbf{X}_i, \mathbf{C})$ becomes maximal [17]. The MI can also consider all the classes simultaneously without the need for individual class pair evaluation, which is more effective and efficient. It is a nonparametric measure as it does not utilize the statistics of the classes of interest. It can also handle class data with arbitrary distributions [15]. The logarithm operation in (1) is nondimensional [18] thus the resulting value does not depend on the coordinates chosen. The use of MI for single feature selection is straightforward by selecting the band having the highest mutual information. The best candidate feature $\mathbf{X}_i$ is selected, if

$$I(\mathbf{X}_i, \mathbf{C}) < I(\mathbf{X}_j, \mathbf{C}) \text{ for all } j \neq i, \text{and } i = 1, 2, \ldots, N. \tag{4}$$

For multiple feature selection, the interaction between the features needs to be taken into account. Ideally, the MI between a group of features and the class data is evaluated directly. However, the number of marginal and joint probabilities to be estimated increases exponentially with the dimensionality and therefore, the evaluation of MI has not been successfully solved for high-dimensional data [15]. On the other hand, an exhaustive search involves C_N^K MI calculations to find the best feature subset of size K out of all the N candidate features. A greedy search strategy was developed in [13] to cope with this difficulty together with the adoption of an 1-D MI evaluation to maximize the relevance and minimize redundancy. This method is referred to as MIFS and is defined as follows. Let $\mathbf{S}_k = \{\mathbf{X}_{s1}, \mathbf{X}_{s2}, \ldots, \mathbf{X}_{sk}\}$ be the subset of k features already selected. The selection of the $(k+1)$th feature is based on the following measure:

$$G(\mathbf{X}_i, k) = I(\mathbf{X}_i, \mathbf{C}) - \beta \sum_{\mathbf{X}_{sk} \in \mathbf{S}_k} I(\mathbf{X}_i, \mathbf{X}_{sk}), \quad \mathbf{X}_i \notin \mathbf{S}_k. \tag{5}$$

The first term in (5) is a measure of the relevance between the feature under examination and the class labels and the second term is a measure of redundancy between this candidate feature and each of the already selected features. β is a user defined parameter which regulates the relative importance of the two terms. If the two parts are treated equally, this parameter is $1/k$ to give an average redundancy to all the currently selected features. The mRMR (max relevance and min redundancy) measure proposed in [14] adopted this approach and (5) becomes

$$G(\mathbf{X}_i, k) = I(\mathbf{X}_i, \mathbf{C}) - \frac{1}{k} \sum_{\mathbf{X}_{sk} \in \mathbf{S}_k} I(\mathbf{X}_i, \mathbf{X}_{sk}), \quad \mathbf{X}_i \notin \mathbf{S}_k. \tag{6}$$

The above method for multiple subset feature selection provides a fast sequential search criterion. However, there is a weakness in this approach due to the way MI has been used. Our proposed improvements are presented in the next section.

B. Improved Feature Selection Technique

1) Normalized Mutual Information: It is difficult to use the value given by (1) directly in an absolute sense, as it is affected by the entropy of the two variables and not bounded to [0, 1]. As we can see from the examples given in Fig. 1, both present perfect one to one relationships, however, the MI values are very different. A few methods to normalize mutual information are available to give a new value between 0 and 1, where 0 is the value for the independent case and 1 for the identical or one-to-one relationship case. The method used in this letter takes the

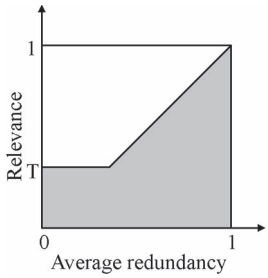

Fig. 2. Constrained searching space.

entropies of both variables into account via geometric mean and is defined as follows [18]–[21]:

$$\widehat{I}(\mathbf{X},\mathbf{C}) = \frac{I(\mathbf{X},\mathbf{C})}{\sqrt{H(\mathbf{X})H(\mathbf{C})}}. \tag{7}$$

The normalized mutual information (nMI) is 1, for both cases in Fig. 1. We introduce nMI for feature selection to make the selection reliable, (6) becomes

$$\widehat{G}(\mathbf{X}_i,k) = \widehat{I}(\mathbf{X}_i,\mathbf{C}) - \frac{1}{k}\sum_{\mathbf{X}_{sk}\in\mathbf{S}_k}\widehat{I}(\mathbf{X}_i,\mathbf{X}_{sk}), \quad \mathbf{X}_i\notin\mathbf{S}_k. \tag{8}$$

The value of both terms is in the range [0, 1]. The candidate feature having the highest difference (high relevance and low average redundancy) is selected as the best $(k+1)^{th}$ feature.

2) Two Constraints: It is possible to select bad features using the method given in (8). The highest difference value may result from two small values, and then the selected features are weakly related to the target. To avoid this problem, the first constraint is introduced

$$\text{if } \widehat{I}(\mathbf{X}_i,\mathbf{C}) < T \text{ remove } \mathbf{X}_i \tag{9}$$

where T is a user-defined threshold, corresponding to the required minimum level of relevance to the target data. It is safer to set T low to include all the relevant features and its value is chosen after examination of the nMI values of the noisy features. This condition can be used as a preprocessing step to remove noisy features before searching the subspace. In this way, the searching time is also significantly reduced. When the highest value of the difference is negative, the selected feature is dissimilar to the already selected features but not highly relevant to the target, which is undesirable. Therefore, the second constraint is given by

$$\widehat{G}(\mathbf{X}_i,k) > 0. \tag{10}$$

If the highest difference value is negative, it means there are no more good features and the largest size of the informative subspace has been reached. This criterion may serves as a stopping rule, if $k < K$. Fig. 2 shows the reduced searching space (unshaded) after the two constraints are imposed.

3) Application of Feature Selection on Principal Components: While the selected features from the original bands keep their spectral meaning, the ignored bands make no contribution to the classification process. To make full use of all spectral measurements, a feature extraction approach is preferred. We propose to apply the feature selection method presented above on principal components [4], [5] to combine feature selection with feature extraction. The new PCA features are given by

$$\mathbf{z} = A^T\mathbf{x} \tag{11}$$

where $\mathbf{x}$ is the original pixel vector and $\mathbf{z}$ is the transformed pixel vector and A^T is the transposed matrix of normalized eigenvectors. For a pixel vector $\mathbf{x}$ located at (p,q) on the image, its new value on the transformed feature i is

$$z_i(p,q) = a_{i1}x_1(p,q) + a_{i2}x_2(p,q) + \ldots\ldots + a_{iN}x_N(p,q) \tag{12}$$

where, a_{ij}, $i,j = 1,2,\ldots,N$ is the element in A. We can see each PCA image is generated from the weighted sum of all the original bands. PCA also has the ability to rearrange the total variance of the data, where the first few PCs contain most of the total variation and the new features are uncorrelated. These useful properties for feature selection will not be offered if a random linear combination of all spectral bands is used. When PCA is applied before feature selection, (8) is modified as

$$\widehat{G}(\mathbf{Z}_i,k) = \widehat{I}(\mathbf{Z}_i,\mathbf{C}) - \frac{1}{k}\sum_{\mathbf{Z}_{sk}\in\mathbf{S}_k}\widehat{I}(\mathbf{Z}_i,\mathbf{Z}_{sk}), \quad \mathbf{Z}_i\notin\mathbf{S}_k. \tag{13}$$

When a subset of principal components is selected in this way, every original spectral band can participate to produce a new feature and hence contribute to the classification task. This is another advantage of the proposed method. We name this method PCA-nMI.

4) Procedure of the Proposed Subspace Detection Approach:

1) Perform PCA on input image $\mathbf{X}$ to obtain $\mathbf{Z}$.
2) Extract training data from each principal component $\mathbf{Z}_i$, and generate class label data $\mathbf{C}$.
3) Evaluate $\widehat{I}(\mathbf{Z}_i,\mathbf{C})$, define threshold T and remove noisy feature $\mathbf{Z}_i$, if $\widehat{I}(\mathbf{Z}_i,\mathbf{C}) < T$.
4) Set the initial selected feature set to the null set, $\mathbf{S}_0 = \{\Phi\}$.
5) Select the 1st feature $\mathbf{Z}_j$, where $\mathbf{Z}_j = \underset{\mathbf{Z}_i}{\operatorname{argmax}}\,\widehat{I}(\mathbf{Z}_i,\mathbf{C})$ and set $\mathbf{S}_1 = \mathbf{S}_0 \cup \mathbf{Z}_j$.
6) Repeat for $k = 1,2\ldots K$, find $\mathbf{Z}_j = \underset{\mathbf{Z}_i\notin\mathbf{S}_k}{\operatorname{argmax}}\,\widehat{G}(\mathbf{Z}_i,k)$. If $\max\limits_{\mathbf{Z}_i\notin\mathbf{S}_k}\widehat{G}(\mathbf{Z}_i,k) > 0$, then $\mathbf{S}_{k+1} = \mathbf{S}_k \cup \mathbf{Z}_j$, otherwise go to step 7.
7) Output $\mathbf{S}_k$ and exit.

III. Experiments

A. Experimental Procedure

The performance of the proposed approach was evaluated using two real hypersepectral image data sets. Image 1 consisted of 191 channels and was collected by the HYDICE sensor over the Washington DC Mall in 1995. Image 2 consisted of 220 channels and was collected by an AVIRIS sensor over the agriculture area of Indian Pine in the north part of Indiana [22]. Sixteen classes are defined in the ground truth [23]. In this experiment, "Grass/Pasture mowed" and "Oats" for the AVIRIS image and "Paths" for the HYDICE were not used as they have insufficient training samples and do not provide

TABLE I
DETAILS OF THE TRAINING AND TEST SAMPLES

HYDICE			AVIRIS		
Class name	Train	Test	Class name	Train	Test
Water	336	392	Grass/trees	96	80
Street	465	367	Soybean-min	130	165
Grass	950	418	Corn-min	15	14
Trees	221	627	Stone-steel	25	20
Roof	121	98	Alfalfa	15	15
Shadow	15	16	Grass/Pasture	108	72
AVIRIS			Corn-notill	48	40
Hay-windrowed	165	135	Soybean-Clean	21	10
Soybean-notill	109	85	Corn-min	64	62
Woods	279	248	Bldg-Grass	20	15
Wheat	42	63			

TABLE II
SELECTED FEATURES

Data	Methods	Order of selected features
HYDICE	mRMR	B:77, 28, 102, 167, 57
	Org-nMI	B: 83, 51, 102, 57, 165
	PCA	PC: 1, 2, 3, 4, 5
	PCA-nMI-WtC	PC: 2, 1, 94, 6, 18
	PCA-nMI	PC: 2, 6, 1, 18, 5
AVIRIS	mRMR	B: 29,1,141,65,16,3,94,168,218,9
	Org-nMI	B: 29,1,141,9,168,70,16,3,199,28
	PCA	PC: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
	PCA-nMI-WtC	PC: 1,3,4,157,2,162,104,5,150,151
	PCA-nMI	PC: 1,3,4,2,9,5,15,16,6,12

overall representative results. The rest are as detailed in Table I. The reference samples are selected randomly based on the ground truth image and then half of the samples are selected for training and the remaining are used for validation purposes. The proposed PCA-nMI method was compared with the following feature reduction methods:

1) mRMR: Feature selection based on a mutual information criteria of max relevance and minimum redundancy defined in (6).
2) Org-nMI: Feature selection from the original bands based on normalized MI defined in (8).
3) PCA: The unsupervised PCA feature extraction method described in Section II-B-3.
4) PCA-nMI-WtC: PCA-nMI without constraints as defined in (13).

B. Feature Extraction Results

The original and PCA images are passed through a quantization process to compute the joint and marginal probability of each feature. The images are quantized into 32 bins in this experiment to avoid very low probability in each bin, unreliable density functions and also for faster processing. The minimum nMI, T, required was set to 0.1 based on the examination of the nMI values of obvious noisy features. For each method, the input features were ranked and the order of the selected features is shown in Table II. The PC-94 for the HYDICE, PC-104 and PC-151 for the AVIRIS data were selected with the PCA-nMI-WtC approach, even though their nMI are 0.096, 0.084 and 0.012, respectively. When the constraints are applied these noisy features $(\mathrm{nMI} < 0.1)$ were not selected. Fig. 3 shows two subspace projections of the HYDICE data illustrating the benefits of applying normalized MI over PCA. While PC-6 has

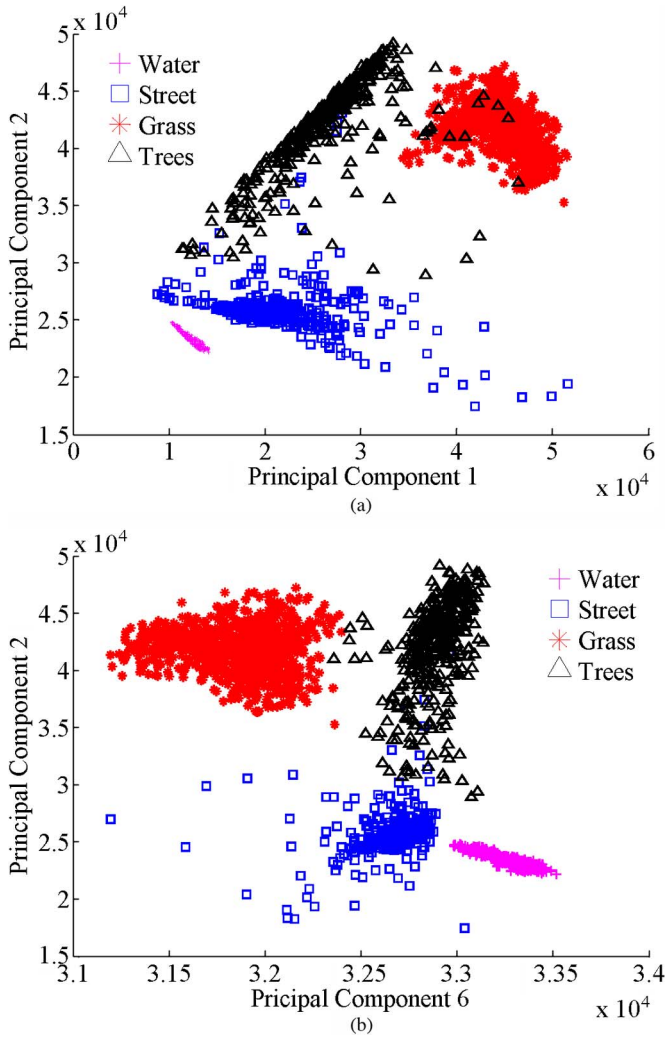

Fig. 3. Feature space of (a) PC-1 and PC-2 and (b) PC-6 and PC-2 for HYDICE data. Better separation among the classes can be observed in (b).

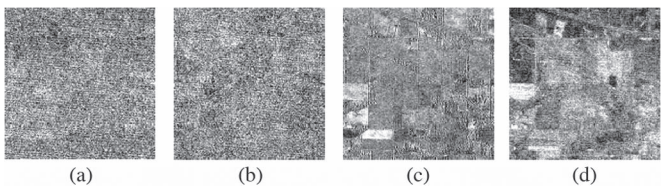

Fig. 4. For the AVIRIS data PCA-nMI selects PC-15 (c) and PC-16 (d) as rank 7 and rank 8 as they are more informative than PC-7 (a) and PC-8 (b).

lower variance than PC-1, the input classes are more separable in the subspace formed by PC-6 and PC-2 than by PC-1 and PC-2. Similarly, PC-7 and PC-8 have higher variance, however, they are not more relevant to the class data than PC-15 and PC-16, which were ranked as 7th and 8th by PCA-nMI as shown in Fig. 4.

C. Classification Results

All the experimental results shown in this section on classification correspond to the average classification accuracy obtained by a support vector machines (SVM) classifier [24]. The SVM Gaussian kernel [25] parameters ($\mathrm{C} = 0.1$ and $\gamma = 0.16$ for HYDICE and $\mathrm{C} = 5$ and $\gamma = 6.08$ for AVIRIS) were selected

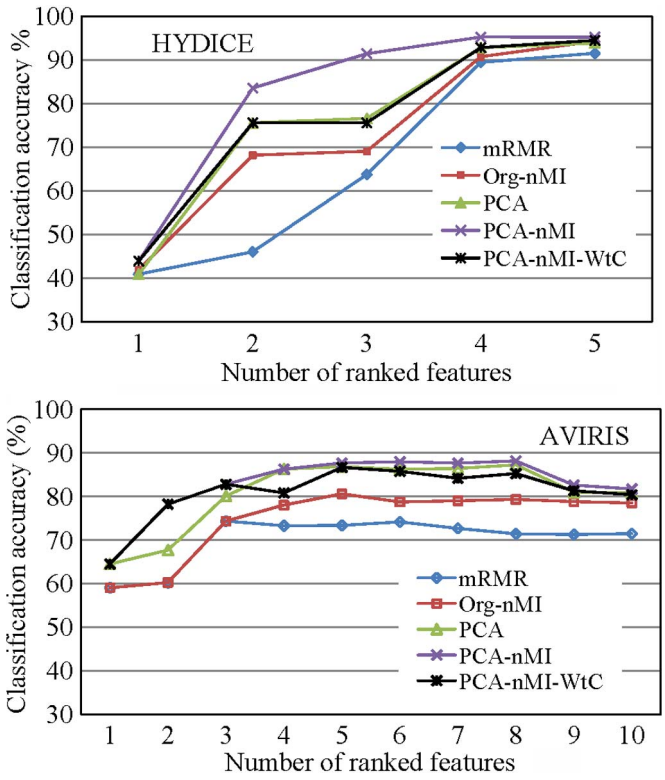

Fig. 5. Classification accuracy of HYDICE and AVIRIS data.

through a 10-fold cross-validation process by using 10 features for AVIRIS and 5 features for HYDICE data. The classification accuracy using the complete set of spectral bands without feature selection/extraction is 66.02% for the HYDICE data and 64.72% for the AVIRIS data. These results are lower than the published results when feature extraction was conducted [26], [27] showing the strong need for feature reduction. For each method and data set, the overall classification accuracies are given in Fig. 5 with respect to the subset of K features selected using the methods investigated in this study. A clear improvement of the classification accuracy has been observed. For the HYDICE data, 4 PCA-nMI features are sufficient to obtain 96% classification accuracy, whereas the other approaches (PCA, mRMR, Org-nMI) require at least five features to obtain classification accuracy close to the proposed approach. However, further improvement may be possible by supplying more training samples and input features. The proposed method also obtains better classification accuracy in almost all cases for the AVIRIS data set.

IV. Conclusion

The feature selection criterion proposed in this study is based on maximizing relevance and minimizing redundancy. It has natural applicability for multiclass problems and non-Gaussian data. The nMI measure offers higher reliability by removing the dependency on each feature's entropy. The two constraints introduced can avoid weak relevant features being selected. The searching space is reduced as well to speed up the process. The reduced number of features given by the proposed method can provide improved results for the cases where only a limited number of training samples are available.

References

[1] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.
[2] X. Jia, B. Kuo, and M. M. Crawford, "Feature mining for hyperspectral image classification," *Proc. IEEE*, vol. 101, no. 3, pp. 676–697, Mar. 2013.
[3] J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis*, 4th ed. Berlin, Germany: Springer-Verlag, 2006.
[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York, NY, USA: Academic, 1990.
[5] L. Ying, G. Yanfeng, and Z. Ye, "Hyperspectral feature extraction using selective PCA based on genetic algorithm with subgroups," in *Proc. ICICIC*, 2006, pp. 652–656.
[6] J. S. Borges and A. R. S. Marcal, "Evaluation of feature extraction and reduction methods for hyperspectral images," in *New Developments and Challenges in Remote Sensing*. Rotterdam, The Netherlands: Millpress, 2007.
[7] M. A. Hossain, M. Pickering, and X. Jia, "Unsupervised feature extraction based on a mutual information for hyperspectral image classification," in *Proc. IGARSS*, Vancouver, 2011, pp. 1720–1723.
[8] F. R. L. Bruzzone 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. 1318–3121, Nov. 1995.
[9] X. Guorong, C. Peiqi, and W. Minhui, "Bhattacharyya distance feature selection," in *Proc. 13th Int. Conf. Pattern Recognit.*, 1996, pp. 195–199.
[10] C. Conese and F. Maselli, "Selection of optimum bands from TM scenes through mutual information analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 48, no. 3, pp. 2–11, Jun. 1993.
[11] C. Cariou, K. Chehdi, and S. L. Moan, "BandClust: An unsupervised band reduction method for hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 565–569, May 2011.
[12] S. Prasad and L. M. Bruce, "Hyperspectral feature space partitioning via mutual information for data fusion," in *Proc. IEEE IGARSS*, 2007, pp. 4846–4849.
[13] R. Battiti, "Using Mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.
[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relecvance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006.
[16] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, pp. 1415–1438, Mar. 2003.
[17] M. L. Murillo and A. A. Rodriguez, "Maximization of mutual information for supervised linear feature extraction," *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1433–1441, Sep. 2007.
[18] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.
[19] A. Strehl and J. Ghosh, "Cluster ensembles a knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Mar. 2002.
[20] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Mar. 2010.
[21] T. O. Kvalseth, "Entropy and correlation: Some comments," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-17, no. 3, pp. 517–519, May 1987.
[22] D. A. Landgrebe. [Online]. Available: https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html
[23] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, Aug. 2009.
[24] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, Apr. 2011.
[25] B. Scholkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2758–2765, Nov. 1997.
[26] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.
[27] S. Aksoy, *Spatial Techniques for Image Classification*, C. H. Chen, Ed., 2nd ed. Boca Raton, FL, USA: CRC Press, 2006, pp. 491–513.