Unsupervised Cluster-based Band Selection for Hyperspectral Image Classification

Jee-Cheng Wu and Gwo-Chyang Tsuei
Department of Civil Engineering, National I-Lan University, I-Lan City 260, Taiwan
wujc@niu.edu.tw, gctsuei@niu.edu.tw

Abstract - A hyperspectral image usually has a large data volume. Several dimensionality reduction (DR) approaches have been investigated to remove redundant information from highly corrected bands. One of the DR approaches is the unsupervised cluster-based band selection (UCBS) method, that is, the bands can be grouped together using different cluster strategies. The method is time consuming, however, because of iterative processing in the band cluster-stage. In this paper, an unsupervised cluster-based band selection method is proposed. The method including two steps is called SensorClust. Firstly the cross-correlation matrix of the entire image was computed and Landsat ETM+ sensor wavelength ranges were used to cluster bands. Secondly in each cluster the covariance matrix was computed, and the bands were selected with the maximum or minimum values along the diagonal of covariance matrix. To demonstrate the effectiveness of the proposed method, a support vector machine (SVM) was selected to carry out supervised classification. The experimental results show the proposed method achieves good classification results in terms of robust clustering.

Index Terms - Hyperspectral image, band selection, dimensionality reduction, Landsat.

I. Introduction

Hyperspectral imagery comprises potentially hundreds of narrow spectral bands, and adjacent bands generally correlate strongly. Removal of redundancy bands can not only save computation time but also improve classification performance. Two techniques reduce hyperspectral dataset dimensionality. First, feature extraction transforms spectral bands into a low-dimensional feature space. Second, band selection selects a subset of spectral bands, which contain most information. Researchers may use prior knowledge (i.e., a supervised manner) to reduce dimensionality and preserve the desired object information. However, the requisite prior knowledge for hyperspectral dataset is usually not available in practice.

Unsupervised feature extraction algorithms (e.g., principal component analysis) are in wide use to reduce hyperspectral dataset dimensionality for classification. Nevertheless, their limitation is that they mathematically transform original bands into feature space, which has no physical meaning in interpretation. Currently, the unsupervised band selection methods still have been proposed to overcome this problem [1][2]. Furthermore, several researchers use different cluster-based strategies to select bands and to reduce dimensionality [3][4][5]. In general, the cluster-based algorithms include two stages: joining similar bands by clustering strategy and selecting the cluster representative band by dissimilarity measure [6]. Nevertheless, cluster-based algorithms are time consuming, due to iterative processing in the band cluster stage.

This paper proposes a novel unsupervised cluster-based band selection method using a fast clustering strategy and covariance matrix to select bands. Section 2 provides a description of the proposed method. Section 3 describes the experimental results. Finally, Section 4 includes some concluding statements and comments on plausible future research.

II. Proposed Unsupervised Sensor Cluster-based Band Selection

Without prior knowledge (e.g., training samples) an unsupervised cluster-based band selection (UCBS) method, called SensorClust, is proposed. The method is not only to reduce hyperspectral dataset dimensionality without prior knowledge (e.g., training samples) but also to cluster bands fast. It includes four steps. Firstly cross-correlation matrix is computed from the entire image. In the cross-correlation matrix, contiguous bands along the diagonal line show in square blocks, representing high correlation among them.

Secondly this paper adopts Landsat ETM+ sensor wavelength ranges; thus grouping the cross-correlation matrix of the hyperspectral dataset into six major clusters. The six major clusters are along the diagonal line with separation lines between blocks at wavelengths: 450 – 515, 525 – 605, 630 – 690, 750 – 900, 1550 – 1750, and 2080 – 2350 nm.

Thirdly the cross-correlation matrix is re-analyzed. If a high correlation block along the diagonal line is not included in the six major clusters, the bands of this block are grouped into an additional cluster.

At last, covariance matrix in each cluster is computed. Along the covariance matrix diagonal, we select the band with the maximum value as the primary band and the band with the minimum value as the secondary band because the two bands have largest dissimilarity measures of the covariance values.

III. Experimental Results

A. Data Sets

This paper uses two hyperspectral images to reduce dimensionality and evaluate classification performance. The AVIRIS – Indian Pine image with 16 classes of labeled samples is available at http://engineering.purdue.edu/~biehl/MultiSpec/ (Figure 1, Left). Researchers randomly selected 15% of the labeled samples from each class for training, using the rest for

validation. The image has 220 contiguous spectral channels covering a spectral region from 400 to 2500 nm in approximately 10 nm bandwidths. In the first scenario, bands 104 – 108, 150–163, and band 220 were not used because of atmospheric water vapor absorption, and only 200 spectral bands were used. In the second scenario, additional bands 1–3, 103, 109 – 112, 148 – 149, 164 – 165, and 217 – 219 were discarded because of low signal to noise ratio (SNR), and only 185 bands were used.

Fig. 1 The two hyperspectral images. (Left) AVIRIS – Indian Pine image, RGB composite (band 30, 20, and 10). (Right) HYDICE – Fort Hood image, RGB composite (band 49, 35, and 18).

The second scene, HYDICE – Fort Hood imagery is available for downloaded at http://www.agc.army.mil/Hypercube/ (Figure 1, Right). The dataset comprises 307 pixels by 307 lines by 210 bands. Researchers generated 11 land use classes and labeled samples for this paper. They randomly selected 5% of each class' labeled samples for training, using the remaining label samples for validation. In the first scenario, bands 104 – 109, 139 – 151, and 206 – 210 were not used due to atmospheric water vapor absorption, and only 186 spectral bands were used. In the second scenario, additional bands 1 – 4, 76, 87, 101 – 103, 110 – 111, 136 – 138, 152 – 153, and 198 – 205 were discarded because of bad quality, and only 162 bands were used [7].

B. Selection of Primary and Secondary Bands

The cross-correlation matrices of the AVIRIS and HYDICE images are computed. The two matrices are blocked at six wavelength ranges: 450 – 515, 525 – 605, 630 – 690, 750 – 900, 1550 – 1750, and 2080 – 2350 nm, and the six major clusters are corresponding to band 7 – 12, 14 – 21, 25 – 32, 39 – 54, 123 – 143, and 178 – 204 at Indian Pine image as shown in Figure 2; and band 16 – 30, 32 – 44, 48 – 54, 60 – 72, 118 – 133, and 164 – 192 at Fort Hood image as shown in Figure 3.

Then, we analyze the two cross-correlation matrices. Following the 750 – 900 nm square blocks, high correlation bands exist between 900 nm and 1352 nm along the matrices' diagonal. This paper assumes two additional clusters, 900 – 1130 nm and 1130 – 1360 nm (see Table I).

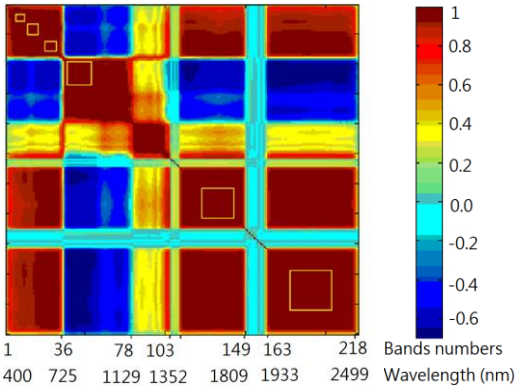

Fig. 2 Cross-correlation matrix of the AVIRIS – Indian Pine, IN image. The diagonal line indicates the highest correlation, 1. Yellow squares represent Landsat ETM+ sensor wavelength ranges.

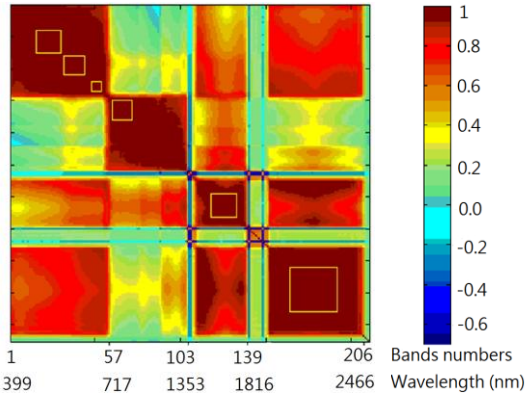

Fig. 3 Cross-correlation matrix of the HYDICE – Fort Hood, TX image. The diagonal line indicates the highest correlation, 1. Yellow squares represent Landsat ETM+ sensor wavelength ranges.

TABLE I Proposed SensorClust for band selection

Types of clusters	wavelength unit: nm
Major clusters	450 – 515/ 525 – 605/ 630 – 690/ 750 – 900/ 1550 – 1750/ 2080 – 2350
Two additional clusters	900 – 1130/ 1130 – 1360

Refer to Table I, in the AVIRIS – Indian Pine image, the primary bands are 12/21/29/42/123/180 from the major clusters, and bands 55/89 from the two additional clusters. The secondary bands are 7/14/31/40/143/204 from the major cluster. In the HYDICE – Fort Hood image, the primary bands are 28/44/53/60/118/168 from the major clusters, and bands 76/90 from the two additional clusters. The secondary bands are 16/34/48/72/127/177 from the major clusters.

C. Parameters used in dimensionality reduction and classifier

To demonstrate the effectiveness of the proposed method, results are compared with a feature extraction method (i.e. principal component analysis, PCA) and two clustering based methods (i.e. hierarchical clustering WaLuMI [8], and recursive binary band-splitting BandClust [9]) in terms of use/cover classification accuracy and Kappa coefficient.

The PCA transformed bands into feature dimensions with dimensions set to 6, 8, 14, 20, and 25. For band selection, the proposed SensorClust clustered bands and selected 6, 8, and 14 bands. The WaLuMI method clustered bands and selected 6, 8, 14, and 20 bands. The BandClust method clustered bands and selected 8 and 15 bands for AVIRIS – Indian Pine image, and selected 7 bands for HYDICE – Fort Hood image (see Table II).

TABLE II The BandClust for Band Selection

Images	Selected bands
AVIRIS - Indian Pine (8 bands)	average(band 4- 28), average(band 28- 41), average(band 41- 76), average(band 76- 99), average(band 99- 102), average(band 113- 134), average(band 134- 140), average(band 140- 147, 166- 185)
AVIRIS - Indian Pine (15 bands)	average(band 4- 28), average(band 28- 32), average(band 32- 36), average(band 36- 41), average(band 41- 47), average(band 47- 54), average(band 54- 76), average(band 76- 99), average(band 99- 102), average(band 113- 115), average(band 115- 122), average(band 122-134), average(band 134-140), average(band 140- 147), average(band 166- 185)
HYDICE – Fort Hood (7 bands)	average(band 1- 8), average(band 8- 30), average(band 30- 40), average(band 40- 55), average(band 55- 103), average(band 110- 138), average(band 152- 205)

Environment for Visualizing Images (ENVI®) [10] performed supervised classification to derive land use/cover classes in the hyperspectral images. We set three parameters (kernel type, gamma in kernel function and penalty) for the support vector machine classifier. The radial basis function (RBF) of kernel type was chosen, the default value of the gamma parameter was used, and the penalty values were set to 100, 1000, and 10000.

D. Results

The scenarios used three random replications to guarantee classification accuracy stability. The best classification results were always obtained using the RBF kernel and the penalty value of 10000; this paper only reports the results obtained using this configuration at each setting dimension. The classification performance of each method is evaluated as the average of kappa coefficients shown in Figure 4 and Figure 5, and we have the flowing findings.

1. If the bands (i.e. water vapor bands, noise bands, and bad bands) are carefully removed from original dataset, we can get the highest kappa coefficient classification accuracy.
2. Classification results show that the PCA feature extraction method outperformed the cluster-based band selection methods with the higher kappa coefficient. However, the PCA method is sensitive to noise. When transforming image spectral bands from spectral space to feature space, we carefully select bands from the dataset to better preserve the feature information.
3. The curse of dimension was already observed in Figure 4 and Figure 5 using the number of the PCA bands larger than 20. Moreover, in Figure 5 the curse of dimension occurred using the number of the WaLuMI bands larger than 14.
4. For unsupervised cluster-based band selection, in Figure 4 the BandClust yields the highest kappa coefficient value with lesser bands. However, the proposed method and the WaLuMI method result in a higher kappa coefficient than the BandClust in Figure 5 from the number of bands larger than 8.
5. The classification results of the proposed method can be compared with the WaLuMI method when the numbers of dimensions increase from 6 to 14 in the two Figures.

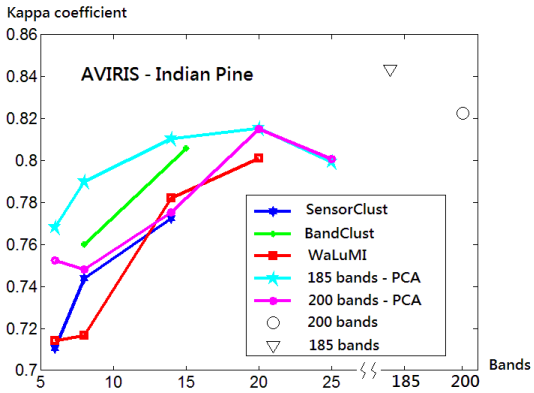

Fig. 4 Classification results obtained on the AVIRIS – Indian Pine image: Effect of dimensionality reduction on kappa coefficient as a function of dimension.

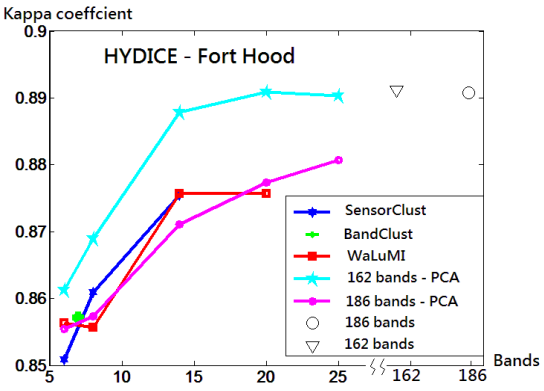

Fig. 5 Classification results obtained on the HYDICE – Fort Hood image: Effect of dimensionality reduction on kappa coefficient as a function of dimension.

IV. Conclusion

The benefits of the proposed method are that the bands are clustered easily and that the selected bands can be interpreted with the physical meanings. Prior knowledge of the satellite sensor wavelength range settings can help us to discriminate spectral information among various features on the Earth. This paper proposes a novel unsupervised cluster-based band selection method using Landsat ETM+ sensor wavelength ranges to cluster bands and reduce data volume. Results are found to be encouraging when the proposed methodology is compared with the two well-known unsupervised cluster-based band selection methods in reducing dimensions for classification. Future research will not only to discover the other implicit clusters from cross-correlation matrix, but also to explore the other satellite sensor wavelength ranges for different hyperspectral image applications, such as geology end-member extraction application.

Acknowledgment

The authors would like to thank Dr. A. Martinez-Uso at the Jaume I University, Castellon de la Plana, Spain, for making the WaLuMI execution code available, and also Claude Cariou at the Laboratoire De Traitement des Signaux et Images Multicomposantes et Multimodales, Ecole Nationale Superieure des Sciences Appliquees et de Technologie, Universite de Rennes 1, Lannion, France, for providing the different band numbers were used in AVIRIS – Indian Pine and HYDICE – Fort Hood images.

References

[1] H. Yang, Q. Du, and G. Chen, "Unsupervised hyperspectral band selection using graphics processing units," IEEE Journal of selected topics in applied earth observations and remote sensing, vol. 4, no. 3, pp. 660-668, September 2011.

[2] S. Jia, Z. Ji, Y. Qian and L. Shen, "Unsupervised band selection for hyperspectral imagery classification without manual band removal," IEEE Journal of selected topics in applied earth observations and remote sensing, vol. 5, no. 2, pp. 531-543, April 2012.

[3] B. Xu. and P. Gong, "Land-use/land-cover classification with multispectral and hyperspectral EO-1 data," Photogrammetric Engineering and Remote Sensing, vol. 73, no. 8, pp. 955-965, August 2007.

[4] Q. Du and H. Yang, "Similarity-based unsupervised band selection for hyperspectral image analysis," IEEE Transactions on Geoscience and Remote Sensing letters, vol. 5, no. 4, pp. 564-568, Oct. 2008.

[5] H. Su, H. Yang, Q. Du, and Y. Sheng, "Semisupervised band clustering for dimensionality reduction of hyperspectral imagery," IEEE Transactions on Geoscience and Remote Sensing letters, vol. 8, no. 6, pp. 1135-1139, Nov. 2011.

[6] P. Bajcsy and P. Groves, "Methodology for hyperspectral band selection," Photogrammetric Engineering and Remote Sensing, vol. 70, .no. 7, pp. 793-802, July 2004.

[7] HyperCube Pictorial User's Guide. 2012. http://www.agc.army.mil/Missions/Hypercube.aspx, (last date accessed: 19 April 2013).

[8] A. Martinez-Uso, F. Pla, J. M. Sotoca, and P. Garcia-Sevilla, "Clustering-based hyperspectral band selection using information measures," IEEE Transactions on Geoscience and Remote Sensing, vol. 45, no. 12, pp. 4158-4171, 2007.

[9] C. Cariou, K. Chehdi, and S. L. Moan, "BandClust: an unsupervised band reduction method for hyperspectral remote sensing," IEEE Transactions on Geoscience and Remote Sensing letters, vol. 8, no. 3, pp. 565-569, 2011.

[10] ENVI 4.6 User's Guide., ITT Visual Information Solutions, Boulder, CO, 2008.