

# Mutual Information-Based Hierarchical Band Selection Approach for Hyperspectral Images

Sonia Sarmah and Sanjib Kumar Kalita

**Abstract** Hyperspectral images consist of hundreds of spectral bands with relatively narrow bandwidth and hence records detailed information of the objects. Due to this detailed and enormous amount of information content, the use of hyperspectral images has become very popular in various fields such as land cover monitoring, agriculture, defense, etc. However, this increased spectral dimension results in increased computational complexity. Hence, the selection of minimal subset of spectral bands to represent the actual information effectively without much degradation is a challenge in the field of hyperspectral image analysis. This paper proposes a hierarchical band selection approach by constructing a spectral partition tree-based on mutual information. Initially, each spectral band has been considered as a leaf node. To minimize the redundancy of information carried by neighboring bands, in every level, new nodes are created by merging adjacent bands or group of bands, for which mutual information has been used as the deciding criterion. Finally from each group of bands, a representative band is selected which jointly form the set of selected bands. Experiment is carried out on the AVIRIS Indian Pines dataset by designing training and testing samples containing only the selected set of bands. The experimental results of the proposed method are found to be very promising and competitive with the existing techniques.

**Keywords:** Hyperspectral images · Spectral partition tree · Entropy  
Mutual information · Support vector machine

---

S. Sarmah (✉) · S.K. Kalita  
Department of Computer Science, Gauhati University, Guwahati, India  
e-mail: sarmahsonia07@gmail.com

S.K. Kalita  
e-mail: sanjib959@rediffmail.com

## 1 Introduction

Hyperspectral sensors capture images with narrow and contiguous spectral bands covering spectrum not only from the visible range, but also from the ultra violet and infra red region. The resultant datasets are three-dimensional which are represented as data cubes of size  $P \times Q \times N$ , where  $P$  and  $Q$  are the spatial dimensions and  $N$  is the spectral dimension. Each spatial plane can be viewed as collection of two-dimensional scenes containing  $P \times Q$  pixels, each taken at a specific wavelength  $\lambda$  and each pixel can be viewed as a vector consisting of  $N$  reflectance values. One such sensor is the airborne visible/infrared imaging spectrometer (AVIRIS) that captures images with up to 224 spectral bands ranging from 400 to 2500 nm [1]. With such high spectral dimension, much more detailed and discriminative information of the objects can be acquired which results in increased classification accuracy. However, performance of many supervised classification methods get strongly affected by increased dimensionality. Hence, spectral dimension reduction is a crucial step in hyperspectral image analysis. Though the hyperspectral images capture a wide range of spectrum, not all bands are equally important for classification purpose as each band does not contribute equally toward the discrimination of the objects. Some bands contain very little or no relevant information. Moreover, the adjacent bands often share redundant information. Thus, this is important to remove such redundant and noisy information before going to the classification phase.

In literature, many dimension reduction techniques are available. Techniques like PCA and ICA map the higher dimensional feature space in to lower dimensional feature space by carrying out linear or nonlinear transformations [2, 3]. In such transformations, original interpretation of spectral data gets compromised. Feature selection is another method for dimension reduction where only the most salient features are selected. This type of methods can broadly be divided into two categories (1) wrappers and (2) filters [4]. In wrappers, a learning algorithm is applied to examine the utility of the features of the dataset. As feature selection is highly dependent on the learning algorithm to be applied, in wrappers it is difficult to switch among the learning algorithms. Moreover, for large databases containing many features, wrappers may become intractable as each and every potential subset of features is evaluated by the learning algorithm. In filters, subset of features is selected as a preprocessing step irrespective of the learning algorithm to be applied and hence is relatively more general compared to wrappers [5, 6].

In feature selection techniques, the main objective is to select the set of features which carry maximum information about the class and yet with minimum correlation between each other. In [2, 7], authors have shown that the adjacent bands of the hyperspectral images are highly correlated and they have utilized the correlation effectively to reduce the number of bands. However in these schemes, the correlation with the class was not considered. In our proposed algorithm, to avoid redundant information carried by adjacent bands, initial band grouping phase is carried out. In this phase, bands are partitioned into  $K$  ( $K \leq N$ ) groups by

constructing a bottom up spectral partition tree by recursively merging adjacent spectral bands with redundant information. In the second and final phase, a band is selected from each group as representative which jointly forms the filtered set of selected bands. Both the phases are discussed in Sect. 3.

## 2 Entropy and Mutual Information-Based Band Selection Scheme

### 2.1 Entropy

In information theory, entropy is the measure of information content of a random variable in terms of uncertainty. Let  $X$  be a discrete random variable with probability distributions  $p(x)$ , where  $x \in X$ , the entropy is defined by

$$H(X) = - \sum_{x \in X} p(x) \log p(x) \quad (1)$$

In literature, entropy has been used as band selection criteria for hyperspectral images [2, 3, 8, 9]. Entropy of each spectral band is calculated to measure their information content. The one with the higher entropy values are selected for classification purpose. As only the bands with highest information contents are selected, the classification accuracy is often high. However, the measured information content by entropy suffers from lack of reference or objective. The information content in some bands, though high, may not have any relevance to the target classification.

### 2.2 Mutual Information (MI)

Mutual information is another technique used to evaluate the effectiveness of a spectral band. Unlike entropy, in mutual information-based approaches apart from measuring the information content of the individual spectral bands, the relevance of the information with the target (reference) image is also considered [10, 11]. In information theory, MI measures the mutual dependence between two random variables. Mathematically, if  $X$  and  $Y$  are two discrete random variables with probability distributions  $p(x)$  and  $p(y)$ , joint probability distribution  $p(x, y)$ , where  $x \in X$  and  $y \in Y$ , then MI is calculated as [12],

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (2)$$

$$= H(X) - H(X|Y) \tag{3}$$

$$= H(X) + H(Y) - H(X, Y) \tag{4}$$

Here,  $H(X)$  and  $H(Y)$  are the entropies of  $X$  and  $Y$ , respectively.  $H(X|Y)$  is the conditional entropy of  $X$  with respect to  $Y$  and  $H(X, Y)$  is the joint entropy of  $X$  and  $Y$ . Equations (5, 6) represents conditional and joint entropy.

$$H(X|Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x|y) \tag{5}$$

$$H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y) \tag{6}$$

In [11], authors have calculated MI of each band in the dataset with the corresponding reference image to measure the effectiveness of the spectral bands in terms of information content about the various classes (Fig. 1).

### 3 Proposed MI-based Hierarchical Band Selection Approach

In this paper, we have proposed a hierarchical band selection scheme based on mutual information. The algorithm is divided into two phases: (i) band grouping and (ii) band selection. Figure 2 shows the proposed MI-based band grouping and selection method.

#### 3.1 Band Grouping Phase

The binary partition tree (BPT) in hyperspectral images was introduced in [12, 13] to merge regions based on spatial features. Inspired by the BPT concept, in the proposed algorithm we have constructed a spectral partition tree (SPT) by iterative bottom up merging of the spectral bands. The individual bands form the initial

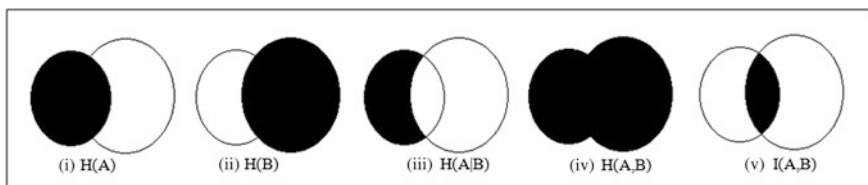


Fig. 1 Diagrammatic representation of entropy and mutual information

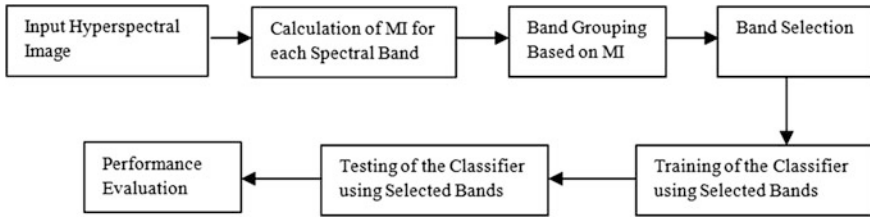
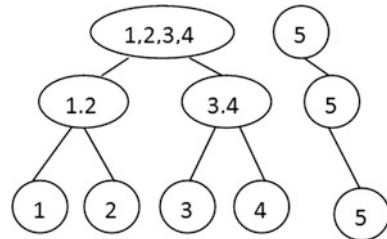


Fig. 2 Block diagram of the proposed method

Fig. 3 Example of construction and representation of the spectral binary tree



nodes, i.e., the leaf nodes of the SPT. Thereafter, in each iteration the nodes for the upper level are created by merging two adjacent nodes. The merging criterion is decided by the difference of MI values of the spectral bands with the reference image and the complementary threshold  $T$ . In the first iteration where the input is the set of  $N$  spectral bands of the image, two adjacent bands  $n$  and  $n + 1$ , are merged if the difference  $d(n) = |\text{MI}(n) - \text{MI}(n + 1)| < T$ . This condition ensures that the adjacent bands which do not carry any significant complementary information are grouped. The new node which is formed as a result of merging is represented in the next level as a set of the merged spectral bands. The nodes which cannot be merged are left as it is for the next iteration. For the rest of the iterations, any two adjacent nodes, which may contain more than one spectral band (as a result of merging in the previous iterations), are merged if  $d_{\text{max}}(n) = |\max(\text{MI}(n)) - \max(\text{MI}(n + 1))| < T$  and  $d_{\text{min}}(n) = |\min(\text{MI}(n)) - \min(\text{MI}(n + 1))| < T$ . Here  $d_{\text{max}}(n)$  is the difference between the largest MI values of the  $n$ th and  $n + 1$  th node respectively. Similarly,  $d_{\text{min}}(n)$  is the difference between the smallest MI values of the  $n$ th and  $n + 1$  th node, respectively. The algorithm stops if two consecutive iterations produce the same set of nodes. Figure 3 depicts an example of construction and representation of spectral binary tree.

### 3.2 Band Selection Phase

In this phase, a representative band from each group of the  $K$  groups formed in band grouping phase is selected. From each group, the band yielding the highest MI value is selected as the group representative. Any representative band having less MI value than a desired threshold  $\theta$ , is discarded.

**Algorithm: Band Selection****Input: band groups formed in phase 1** $i \leftarrow 1$  $S \leftarrow \emptyset$ **while**  $i \leq K$  **do**select the band B, with maximum MI value from the  $i$ th group.**if**  $MI(B) < \theta$  **then do** $S \leftarrow S \cup B$ **else** $S \leftarrow S \cup S$ **end if****end while**

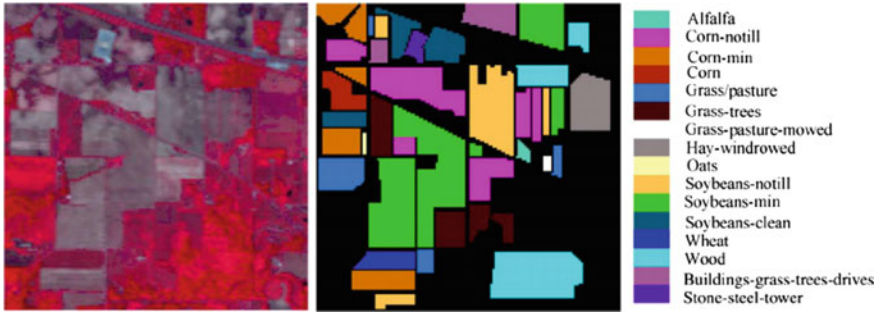
## 4 Experimental Setup and Results

### 4.1 Dataset Description

For our experiment, we have used the AVIRIS Indian Pines Dataset which was collected over Northwest Indiana in June 1992 by AVIRIS sensor. The dataset contains the Indian pine image consisting of 220 spectral bands and 21,025 pixels ( $145 \times 145$ ). The image is accompanied with a reference image where 10,249 pixels are labeled with a number from 1 to 16, denoting the class to which a pixel belongs to. The rest 10,776 pixels are labeled as 0 indicating that these either belong to areas which are not of interest or could not be labeled due to technical difficulties. Out of the 16 classes only nine classes, as listed in Table 1, have been considered for our experiment. Other classes are discarded as the rest have relatively limited number of labeled samples (Fig. 4).

**Table 1** Numbers of samples for different classes in Indian pine image dataset

S. No.	Class	Number of samples
1	Corn-no-till	1428
2	Corn-min-till	830
3	Grass-Pasture	483
4	Grass-Trees	730
5	Hay-windrowed	478
6	Soybean-no-till	972
7	Soybean-min-till	2455
8	Soybean-clean	593
9	Woods	1265



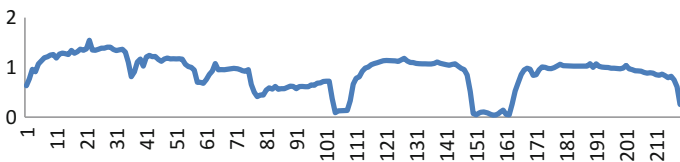
**Fig. 4** False color image of the reference image for Indian pines dataset showing the various classes present

### 4.2 Experiment and Result

Experiment was performed by first calculating the MI of each spectral band of the Indian Pine image with the reference map. Figure 5 shows the calculated results plotted against the band numbers. It can be seen that some band have very low MI values. This is due to water absorption in those bands.

In the band grouping phase, the construction of the spectral binary tree was experimented with different thresholds. For band selection phase, the threshold  $\theta$  was set as 0.30. Number of groups formed in band grouping phase and number of selected representative bands from the groups for each threshold are listed in Table 2. For classification, support vector machine was used as it is known to work well with high input space. For training and testing, 200 random samples from each class with selected bands were taken. The classifier was also tested with samples from the original Indian pines image (containing 220 bands) and the corrected image (containing 200 bands). Table 2 shows the achieved experimental results.

From the experimental results, it is observed that as we limit the number of bands, the overall accuracy (OA) of classification also decreases. However, this decrease is not that significant and results are still comparable with that when all the bands are used. As the number of bands decreases from 200 to 38, the classification accuracy decreases by only 0.23%. However, when the number of bands further gets decreased to 27 the degradation in classification accuracy is a bit higher. The most significant observation is that with all the 220 bands the achieved OA is lower



**Fig. 5** Mutual information of the spectral bands of the Indian Pie image with the reference image

**Table 2** Experimental results of the proposed algorithm

Threshold	No. of groups formed in band grouping phase	No. of representative bands selected	Overall accuracy (%)	Degradation in overall accuracy (%)
None	220 (using all the bands of the dataset)	220	86.55	
None	200 (using the bands of the corrected image)	200	87.06	-0.51
0.10	42	38	86.83	0.23
0.15	31	27	85.67	1.16
0.20	23	20	85.33	0.34
0.25	19	17	84.77	0.56

than that using 38 bands. This is due to the rejection of the groups of bands which have very low MI with the target image. The proposed algorithm discards the bands which are noisy and do not contain any relevant information are avoided. In the corrected image with 200 bands, the noisy bands are already removed and classification accuracy is higher.

## 5 Conclusion

In this paper, we have proposed a hierarchical band selection approach based on mutual information. A spectral partition tree is constructed to group the adjacent bands carrying redundant information. From each group in the band selection phase, only one band which carries maximum information about the target image is selected to eliminate redundancy. From the results, it has been seen that the bands that get selected through the proposed algorithm, though limited, are highly capable of representing original image without losing much information. Through this algorithm, the noisy and irrelevant bands gets automatically discarded in the band selection phase. When the threshold is high, very few numbers of bands get selected and the achieved overall accuracy of classification is also low. As more number of bands is added to the selected set, the overall accuracy also improves significantly. However, after some limit, this growth in classification accuracy with number of selected bands tends to be slower. The additional accuracy achieved becomes insignificant compared to the increase in spectral dimension.

## References

1. AVIRIS: Airborne visible/infrared imaging spectrometer. <http://aviris.jpl.nasa.gov/>
2. Chang, C.-I., Qian, D., Sun, T.-L., Althouse, M.: A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **37**(6), 2631–2641 (1999)



3. Velez-Reyes, M., Jimenez, L.: Subset selection analysis for the reduction of hyperspectral imagery. *IEEE Int. Geosci. Remote Sens. Symp. Proceed.* **3**, 1577–1581 (1998)
4. Kohavi, R., John, G.: Wrappers for feature subset selection. *Artif. Intell.* **97**, 273–324 (1997)
5. Al-Ani, A.: A dependency-based search strategy for feature selection. *Expert Syst. Appl. Int. J.* **36**, 12392–12398 (2009)
6. Bonev, B., Escolano, F., Angel-Cazorla, M.: A novel information theory method for filter feature selection. *MICAI 2007: Advances in Artificial Intelligence*, pp. 431–440. Springer, Berlin (2007)
7. Kumar, S., Ghosh, J., Crawford, M.M.: Best-bases feature extraction algorithms for classification of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **39** (2001)
8. Groves, P., Bajcsy, P.: Methodology for hyperspectral band and classification model selection. In: *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, pp. 120–128 (2003)
9. Bajcsy, P., Groves, P.: Methodology for hyperspectral band selection. *Photogram. Eng. Remote Sens. J.* **70**(7), 793–802 (2004)
10. Conesea, C., Masellia, F.: Selection of optimum bands from TM scenes through mutual information analysis. *ISPRS J. Photogram. Remote Sens.* **48**(3), 2–11 (1993)
11. Guo, B., Gunn, S., Damper, B., Nelson, J.: Adaptive band selection for hyperspectral image fusion using mutual information. In: *7th International Conference on Fusion* (2005)
12. Novovicova, J., Somol, P., Haindl, M.: Conditional mutual information based feature selection for classification task. In: *CIARP 2007*, vol. 4756, pp. 417–426, LNCS (2007)
13. Valero, S., Salembier, P., Chanussot, J.: New hyperspectral data representation using binary partition tree. In: *Proceedings of IEEE IGARSS Conference*, pp. 80–83 (2010)