Report on Band Selection

Adolfo Martínez-Usó

January 10, 2011

Work reference:

Title:	Clustering-based hyperspectral band selection using information measures
Authors:	Adolfo Martínez-Usó, Filiberto Pla, José M. Sotoca, Pedro García-Sevilla
Journal:	IEEE Transactions on Geoscience and Remote Sensing, vol. 45, n. 12
Date:	December 2007

Contents

Abstract

Hyperspectral imaging involves large amounts of information. This work presents a technique for dimensionality reduction to deal with hyperspectral images. The proposed method is based on a hierarchical clustering structure to group bands to minimise the intra-cluster variance and maximise the inter-cluster variance. This aim is pursued using information measures, such as distances based on Mutual Information or Kullback-Leibler divergence, in order to reduce data redundancy and non-useful information among image bands. The technique presented has a stable behaviour for different image data sets, and a noticeable accuracy, mainly when selecting small sets of bands.

1 Introduction

A very desirable pre-processing step in hyperspectral imaging, and particularly on pixel classification tasks, is to perform a band selection process to reduce the redundant information in the image representation without losing classification accuracy in a significant way and using no supervised information. To this end, we propose a new technique that,

- exploits band correlation through a clustering based algorithm. A similar strategy has also been used in distributional clustering for text categorisation [5] or data compression [7], due to the high dimensionality that these approaches have to deal with.
- can use different measures to discriminate among the bands. In this work, two different measures are proposed, resulting on two variants of the same algorithm. Both criteria are based on Information Theory measures [4].
- obtains subsets of relevant bands to try to get the best classification performance, mainly when selecting small sets of bands, where the band selection methods have to really show their capabilities to extract the relevant information in the data.
- it is not a ranking or incremental method. That is, the best m bands could not be the best $m - 1$ bands plus another relevant band.

This report is organised as follows: Section 2 gives a short guide to common usage. Section 3 briefly lays the theoretical foundations of this SW as well as some implementation and computational issues. Finally, in section 4 some conclusions are offered.

2 Short guide for users

This report presents a software that has been called $BandSelection\_TGRS07$. Linux and windows versions are provided and both share this name.

Help : If just this name is prompted, the following help is given:

```
> BandSelection_TGRS07

From the initial input bands, Kfin bands will be selected.
This SW is based on the work published in:
Martinez-Uso A., Pla F., Sotoca J.M., Garcia-Sevilla P.
Clustering-based hyperspectral band selection using information measures
IEEE Transactions on Geoscience and Remote Sensing.
Vol.45(12), Part 2 pp. 4158-4171. December 2007.

method = 1: Runs WaLuMI algorithm, which is based on Mutual Information
         2: Runs WaLuDi algorithm, which is based on K-L divergence
Kini = Initial number of clusters
Kfin = Final number of clusters

Usage: BandSelection_TGRS07 <method> <Kini> <Kfin> <SourceIm1>...<SourceImN>
```

***method* parameter** : This SW implements a hierarchical clustering that applies two criteria. If $method = 1$, it applies a criterion function based on Mutual Information (see section 3.1.1

for details) whereas, if $method = 2$, it applies a criterion function based on KL-divergence (see section 3.1.2 for details).

***K* parameter** : The hierarchical clustering is done in an agglomerative way. Therefore, this parameter selects the final number of clusters in this clustering process. Once the clusters are created, a representant from each one is chosen and, thus,

- *Kini*: parameter that works as the initial number of bands that must be saved in text files.
- *Kfin*: parameter that works as the final number of bands desired at the end of the process.

Finally, it is important to point out that, obviously, $Kini, Kfin \leq$ number of input bands (DIM) and $Kini \geq Kfin$.

source/input bands : List of bands introduced by the user. These bands participate in the selection process. You must provide at least one and they must be *PGM* files in raw format ($P5$). For the windows version, it is highly recommended the use of text files in order to copy/paste the names of the input bands. This is due to the fact that regular expressions, in the way of linux, are not supported in windows.

Error : An error in the input parameters provokes an output with the previous help.

Output I : The application screens the $Kfin$ number of input bands and which ones have been selected.

Output II : The application also saves some files called “clusters_posi_Noutof$D.method$” and “clusters_name_Noutof$D.ext$”. These files will have a list of N bands selected, out of D, where D number would obviously be the dimension of the image (the number of input bands that have been introduced, DIM). N value ranges from $Kini$ to $Kfin$. These files will have the extension (ext) that depends on the *method* parameter, that is, if $method = 1$, $ext = walumi$ whereas, if $method = 2$, $ext = waludi$. Finally, files labelled with “*_posi_*” at their file name contain the *position* of the selected bands in the list of *source/input bands*, being position 0 the first band. In the same way, files labelled with “*_name_*” at their file name contain the *names* of the image bands selected (the particular *PGM* file)[1].

Example : Imagine that you want to know the 2 most relevant bands from a set of 128 input bands ($DIM = 128$) according to the criterion based on mutual information. Let us also suppose that you are interested in knowing the relevant subsets of bands from the step where the algorithm reach 50 clusters. To this end, you should execute the SW in the following way:

```
> BandSelection_TGRS07 1 50 2 band001.pgm band002.pgm ... band128.pgm

From input bands (DIM=128) -> [band004.pgm] [band116.pgm] selected
Clustering time = 8.50 s.
```

[1] Note that the band number could not be the same than the number that indicates the position in the list of input bands, since they could have been introduced out of order to the application.

Obviously all the bands must be prompted (we have used dots instead writing 128 bands for this example!!) but in the linux version regular expressions can be also used. Thus, from an input set of 128 bands, *band004.pgm* and *band116.pgm* have been selected as the most relevant ones and the process finished in 8.5 seconds. In addition, if you check the current directory, you will see that files,

- "*clusters_posi_05outof128.walumi*", "*clusters_posi_06outof128.walumi*", ..., "*clusters_posi_50outof128.walumi*" and
- "*clusters_name_05outof128.walumi*", "*clusters_name_06outof128.walumi*", ..., "*clusters_name_50outof128.walumi*"

have been created with the subsets of selected bands from $K = 5$ to $K = 50$. They receive the "*.walumi*" extension pointing out which methodology has been used.

3 Clustering-based Band Selection

In hyperspectral imaging for remote sensing, it is very common to have very little or no labelled information. Therefore, a band selection technique that uses no supervised information can be really useful. On the other hand, techniques that do not use supervised information can use the whole data set available, while supervised data sets usually provide labelled information of only one part of the available data.

The band selection technique here proposed is based on a clustering process performed in a similarity space defined among bands. Against other techniques that rank bands by means of a similarity measure, a process that joins similar bands together is proposed, constructing a family of derived clusters that preserves a low variance among the bands that belong to the same cluster and a high variance among different clusters, in an analogous way as clustering is used in vector quantization for data compression. The final selected bands will be the best representative instances from each cluster. Moreover, in contrast to other authors that use a divisive clustering approach [5], we advocate for an agglomerative clustering strategy, in order to also reflect the hierarchical nature of the spectrum structure [9].

In addition to these essential requirements, one of the main objectives of this work is a significant reduction of the redundant information, keeping a high accuracy in classification tasks. To this end, from Information Theory [1], we can find information measures that can quantify how much a given random variable can predict another one. We will particularly focus on this property. Therefore, we propose the use of two different measures to exploit this point: the mutual information and the Kullback-Leibler divergence.

Mutual information is not only widely used as a criterion for measuring the degree of independence between random variables, but it also measures how much a certain variable can explain the information content about another variable, being a generalised correlation measure. Thus, a dissimilarity measure between two bands (random variables) can be defined based on this measure as a relevance criterion. On the other hand, the Kullback-Leibler divergence has been employed as a measure of discrepancy between any two probability distributions, and it can be interpreted as the cost of substituting a given probability distribution by another one. This criterion was already applied to compare hyperspectral image bands [2].

3.1 Dissimilarity measures

3.1.1 Mutual Information-based criterion ($method = 1$, WaLuMI algorithm)

The first dissimilarity measure proposed tries to identify the subset of selected bands that are as much independent as possible among them. It is known that independence between bands [9] is one of the key issues to obtain relevant subsets of bands for classification purposes.

The use of information measures, like Mutual Information, in order to quantify the degree of independence, provides a methodology to find generalised correlations among image bands. Thus, this technique exploits this concept for band selection in order to reduce data redundancy and non-useful information.

Let us introduce some Information Theory concepts and properties [4][12]. The Shannon entropy of a random variable X with probability density function $p(x)$ for all possible events $x \in \Omega$ is defined as,

$$H(X) = -\int_{\Omega} p(x) \ \ \log p(x) \ \ dx \tag{1}$$

In the case of a discrete random variable X, entropy $H(X)$ is expressed as,

$$H(X) = -\sum_{x \in \Omega} p(x) \ \ \log p(x) \tag{2}$$

where $p(x)$ represents the mass probability of an event $x \in \Omega$ from a finite set of possible values. Entropy is often taken as the related amount of *information* of a random variable.

On the other hand, *mutual information* (I) is a measure of independence between random variables. I can be interpreted as a generalised correlation measure, which includes the linear and non-linear dependence between variables. In other words, mutual information quantifies the statistical dependence of random variables, or how much a variable can predict another one.

Due to the complexity in calculating the joint distribution in high dimensional spaces [4], estimation of $I(\hat{\mathbf{S}}, \mathbf{S})$, where $\hat{\mathbf{S}}$ is a subset of random variables out of the original set $\mathbf{S}$ such that $\hat{\mathbf{S}} \subset \mathbf{S}$, becomes complex and highly computational expensive. This is a critical issue from a practical point of view. In this sense, the technique here proposed tries to overcome this drawback by only using comparisons between pairs of random variables, through defining a similarity measure based on the mutual information between two random variables.

Let us consider a set of L random variables that represent their corresponding bands $X_1, ..., X_L$ from a hyperspectral image. $I(X_i, X_j)$ is defined as,

$$I(X_i, X_j) = \sum_{x_i \in \Omega} \sum_{x_j \in \Omega} p(x_i, x_j) \ \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \tag{3}$$

I is always a non negative quantity for two random variables, being zero when the variables are statistically independent. The higher the I, the higher the dependence between the variables is. Furthermore, the following property about two random variables always holds:

$$0 \ \leq I(X_i, X_j) \ \leq \ min\{H(X_i), H(X_j)\} \tag{4}$$

Mutual information I can be expressed in terms of entropy measures according to the following expression:

$$I(X_i, X_j) = H(X_i) + H(X_j) - H(X_i, X_j) \tag{5}$$

where $H(X_i, X_j)$ is the joint entropy; which is defined from the joint probability distribution $p(x_i, x_j)$.

So far, I has been introduced as an absolute measure of common information shared between two random sources. However, as we can infer from equation (5), I by itself would not be suitable as a similarity measure. The reason is that it can be low because either the X_i, X_j variables present a weak relation (such as it should be desirable) or because the entropies of these variables are small (in such a case, the variables contribute with little information). Thus, it is convenient to define a proper measure so that it works independently from the marginal entropies and also measures the statistical dependence as a similarity measure.

Thus, the following measure of similarity between two random variables will be used,

$$NI(X_i, X_j) = \frac{2 \cdot I(X_i, X_j)}{H(X_i) + H(X_j)} \tag{6}$$

which is a normalised measure of I. Furthermore, this normalised mutual information is used as a dissimilarity or distance measure as follows [6]:

$$D_{NI}(X_i, X_j) = \left(1 - \sqrt{NI(X_i, X_j)}\right)^2 \tag{7}$$

3.1.2 Divergence-based criterion ($method = 2$, WaLuDi algorithm)

Another information measure to be considered is the Kullback-Leibler divergence, which can be interpreted as a kind of dissimilarity distance between two probability distributions, though it is not a real distance measure because it is not symmetric. Thus, a symmetric version of the Kullback-Leibler divergence is often used [4][12].

Let us call X_i and X_j two random variables defined in a Ω space, representing the ith and jth bands of a hyperspectral image. Let us assume that $p_i(x)$ and $p_j(x)$ are the probability distributions of these random variables. Thus, the symmetric Kullback-Leibler divergence can be expressed in the discrete domain as follows:

$$D_{KL}(X_i, X_j) = \sum_{x \in \Omega} p_i(x) log \frac{p_i(x)}{p_j(x)} + \sum_{x \in \Omega} p_j(x) log \frac{p_j(x)}{p_i(x)} \tag{8}$$

The Kullback-Leibler divergence is always non-negative, being zero when $p_i(x)$ and $p_j(x)$ are the same probability distribution. This divergence measure can be used as a criterion to know how far two distributions are, and it can be interpreted as the cost of using one of the distributions instead of the other one. In the hyperspectral band selection framework, it can be used as a measure of dissimilarity between two image bands, represented by their corresponding probability distributions.

This divergence measure is the second criterion proposed to be used as a distance for the clustering process, and it has been frequently used in order to compare different probability distributions, also in hyperspectral imaging to measure the overlapped information contained in a pair of image bands, as a band-decorrelation algorithm [2].

3.2 Variance-Reduction Clustering Strategy

Using the introduced criteria either based on the mutual information or the Kullback-Leibler divergence as a dissimilarity measure between two image bands, a hierarchical clustering process is then proposed, in order to form clusters of bands as similar as possible among them within each cluster. The clustering is part of a information compression process, and, at the end of the clustering process, a representative band for each cluster is selected, which will substitute all bands in the cluster, at the lowest possible cost in terms of information loss. The selected representatives will constitute the subset of bands selected, as a compressed image band representation for the whole original set of image bands.

3.2.1 Hierarchical clustering

Hierarchical structures are a very intuitive way to summarise certain types of data sets. One interesting characteristic of hierarchical methods is the fact that different linkage strategies create different tree structures. The algorithm here proposed uses an agglomerative strategy. Thus, the number of groups is reduced one by one.

In particular, a hierarchical clustering algorithm based on a Ward's linkage method [11] is used. Ward's linkage has the property of producing minimum variance partitions. Thus, this method is also called minimum variance clustering, because it pursues to form each possible group in a manner that minimises the loss associated with each grouping (internal cohesion). Several studies point out that this method outperforms other hierarchical clustering methods [8], but, in our case, the process also helps us to form groups with low variance in their level of similarity.

Briefly summarising the linkage strategy, let us suppose that clusters C_r and C_s are merged. The general expression for the distance between the new cluster (C_r, C_s) and any other cluster (C_k) is defined as:

$$\begin{aligned} D[(C_k),(C_r,C_s)] = \alpha \cdot D(C_k,C_r) + \beta \cdot D(C_k,C_s) + \\ + \ \gamma \cdot D(C_r,C_s) + \delta \cdot |D(C_k,C_r) - D(C_k,C_s)| \end{aligned} \tag{9}$$

where α, β, γ and δ are the merging coefficients. Ward's inter-cluster distance results from the following coefficients,

$$\alpha = \frac{n_r + n_k}{n_r + n_s + n_k}, \quad \beta = \frac{n_s + n_k}{n_r + n_s + n_k},$$
$$\gamma = \frac{-n_k}{n_r + n_s + n_k}, \quad \delta = \emptyset,$$

where n_i is the number of instances in group i.

The algorithm starts with the disjoint partition where each cluster is formed as a single pattern (hyperspectral band). At this step, the dissimilarity matrix $D_{L\times L}$ is initialised by means of either of the dissimilarity measures described on sections 3.1.1/3.1.2. After that, the algorithm looks for the two most similar clusters that will have the minimum distance value in matrix $D_{L\times L}$. Then, these two clusters are merged into one and matrix $D_{L\times L}$ is updated using expression (9). The rows/columns corresponding to the merged clusters are deleted and a row/column for the new cluster is added.

This process is repeated until the K number of desired clusters are obtained. The resulting mutually exclusive clusters represent groups of highly correlated bands, and bands from two different clusters will have low correlation.

3.2.2 Selecting cluster representatives

Let us consider now a resulting cluster C with R bands. A weight of each band $X_i \in C$ is defined as,

$$W_i = \frac{1}{R} \sum_{j \in C, j \neq i} \frac{1}{\epsilon + D(X_i, X_j)^2} \tag{10}$$

where ϵ is a very small positive value to avoid singular values, and function $D(X_i, X_j)$ returns the distance value between bands i,j. The representative band from each group is selected as the band with the highest W_i in the cluster.

A low value of W_i means that the band i has an average large distance from the other bands in the cluster, that is, in this case, the band i will have an average low correlation with regard to the other bands in the cluster. In a reverse way, a high value of W_i means that band i has, in average, a high correlation with regard to the other bands in the cluster.

Therefore, when selecting cluster representative bands by using dissimilarity measure D_{NI}, choosing the band in the cluster with the highest average correlation (mutual information) with regard to the other bands in the cluster, it is equivalent to choose the band that better predicts the information content of the other bands in the cluster, since the more mutual information two random variables share, the more can predict one of the variable about the other one and, in this sense, having a high degree of dependence among them.

On the other hand, when selecting cluster representative bands by using distance D_{KL}, choosing the band in the cluster with the highest average divergence with regard to the other bands in the cluster, is equivalent to select the band that would produce the lowest cost, in the average sense, when substituting every band in the cluster by its representative.

As a result of the algorithm, there will be K bands selected, representing K different clusters. The bands within the same cluster will have a high correlation. The selected bands will also cover the dissimilarity space, being a compressed representation that tries to explain most of the information contained in the original representation.

3.3 Implementation and computational issues

One of the key implementation issues is the estimation of the probability function $p(x)$ for each band X. The probability density function $p(x)$ for all event $x \in \Omega$, where Ω is the set of possible values a random variable X can take, will be estimated for each image band as $p(x) = \frac{h(x)}{MN}$, being $h(x)$ the gray level histogram and MN a normalising factor, which is the number of pixels in the image. In an analogous way, to estimate the joint probability distribution $p(x_i, x_j)$ between two bands X_i and X_j, the corresponding joint histogram $h(x_i, x_j)$ is first required and the probability distribution is then computed as $p(x_i, x_j) = \frac{h(x_i, x_j)}{MN}$.

The whole algorithm can be divided into two main parts: the operations done before the clustering (*pre-clustering*), and the operations properly involved in the hierarchical clustering process (*clustering*). Note that, in the *pre-clustering* part, we shall distinguish two different processes depending on the measure used when we calculate the distances between any pair of bands: based on mutual information (D_{NI}) or based on the divergence criterion (D_{KL}).

- **Pre-clustering**: When the process begins, each band in the image is considered as a separated cluster. Then, a distance matrix of size $L \times L$ is initialised with the corresponding distances between pairs of bands, obtaining a symmetrical matrix:

  - When using the distance based on mutual information (D_{NI}), the histogram for each single band and the co-joint histogram for each pair of bands must be computed . Thus, assuming that $MN > G$, where G is the number of gray levels in the bands, the temporal cost of this part is $O(L^2 MN)$.
  - When using the divergence criterion (D_{KL}), the co-joint histograms are no required. Now, the temporal cost of this part is $O(LMN + L^2 G)$.

  Although the D_{KL} criterion requires less computational effort for the matrix initialisation, both methods are computationally affordable. From the point of view of the spatial cost, only the distance matrix, the histograms and the image bands are required in the process. Furthermore, only one pair of bands is required in main memory at a time when the co-joint histograms must be computed.

- **Clustering**: This part is related to the operations that the Ward's linkage method involves. Once the distance matrix has been initialised, its minimum value is found in order to choose two clusters to be joined. Then, a new row and a new column are added for the new cluster created and their entries in the distance matrix are computed according to equation (9). Afterwards, the rows and columns for the old clusters in the distance matrix are removed. The process repeats until the desired number of clusters is reached. There are no additional requirements of memory in this step. The temporal cost of this part is $O(L^3)$ which will be significantly lower than the pre-clustering part for usual images. Only if we had to deal with small images and a large number of bands, the temporal cost of this part would be comparable with the cost of the pre-clustering part.

As an illustrative example of these computational costs, using an Intel Pentium 4 CPU 3.00GHz and the HyMap image with 128 bands of 700 × 670 pixels (see [10] for a database description), our current implementation required about 47 seconds for the *pre-clustering* part (using D_{KL}), whereas the *clustering* part required just 2 seconds. Table 1 gives a simple quantitative analysis of the computational cost for several band selection methods for HyMap image. Note that no optimisations were considered but a direct implementation of the algorithms described were run. Implementations were developed in C++. For LCMV-CBS methods, it is important to point out that the practical optimisation described in [3] has been assumed, that is, since BCM, BDM and BCC, BDC share the resulting bands selected, the best time of each pair has been taken into account, otherwise BDM and BDC methods are much more time-consuming.

Summarising, the method here presented is computationally affordable, even for hyperspectral images with a large number of input bands, since it is based on probability estimations from histogram pixel values of, at most, pairs of bands, avoiding unfeasible high dimensional probability estimations.

Table 1: Processing times for several band selection methods using the HyMap image with 128 bands [10]

	WaLuMI	WaLuDi	BCM-BDM	BCC-BDC	MVPCA	ID
CPU time	1m48s	49s	141m58s	149m47s	3m10s	10s

4 Conclusions

A technique for band selection in multi/hyper-spectral images has been reported, aimed at removing redundant information while keeping significant information for further classification tasks. The proposed method uses a clustering process strategy to group bands that minimises the intra-cluster variance and maximises the inter-cluster variance, in terms of some dissimilarity measures based on band information content, like mutual information and Kullback-Leibler divergence.

The results obtained, from the point of view of pixel classification in hyperspectral images, provide experimental evidence about the importance the proposed clustering strategy in a band dissimilarity space plays in the problem of classification.

The band selection method here presented is fully unsupervised, that is, it uses no labelling information. It is computationally affordable since it is based on probability estimations from histogram pixel values, as a maximum, from pairs of bands, avoiding unfeasible high dimensional probability estimations.

Contact us

Adolfo Martínez-Usó	auso@uji.es
Filiberto Pla	pla@uji.es
José M. Sotoca	sotoca@uji.es
Pedro García-Sevilla	pgarcia@uji.es
Web page:	`www.vision.uji.es`
Department:	Lenguajes y Sistemas Informáticos.
University:	University Jaume I. Castellón. Spain.

References

[1] J. Aczel and Z. Daroczy. *On measures of information and their characterization*. Academic Press, New York, 1975.

[2] Chein-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 on Geoscience and Remote Sensing*, 7(6):2631–2641, November 1999.

[3] Chein-I Chang and S. Wang. Constrained band selection for hyperspectral imagery. *IEEE Trans on Geoscience and Remote Sensing*, 44(6):1575–1585, June 2006.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[5] I. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, 2003.

[6] Raquel Dosil, Xosé R. Fdez-Vidal, and Xosé M. Pardo. Dissimilarity measures for visual pattern partitioning. *LNCS*, (3523):287–294, 2005.

[7] Allen Gersho and Robert M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1992.

[8] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[9] S. Kumar, J. Ghosh, and M. M. Crawford. Best-bases feature extraction algorithms for classification of hyperspectral data. *IEEE Trans on Geoscience and Remote Sensing*, 39(7):1368–1379, 2001.

[10] Adolfo Martinez-Uso, Filiberto Pla, José M. Sotoca, and Pedro Garcia-Sevilla. Clustering-based hyperspectral band selection using information measures. *IEEE Trans on GRS*, 45(12):4158–4171, December 2007.

[11] John H. Ward. Hierarchical grouping to optimize an objective function. *American Statistical Association*, 58(301):236–244, 1963.

[12] Andrew Webb. *Statistical Pattern Recognition*. Wiley, 2nd edition edition, 2002.