Clustering-based multispectral band selection using mutual information[*]

Adolfo Martínez-Usó Filiberto Pla Jose M. Sotoca Pedro García-Sevilla
Departamento de Lenguajes y Sistemas Informáticos
Universitat Jaume I, 12071 Castellón (Spain)
{auso, pla, sotoca, pgarcia}@lsi.uji.es

Abstract

This work presents the application of a novel technique on dimensionality reduction to deal with multispectral images. A distance based on mutual information is used to construct a hierarchical clustering structure. Experimental results show that the method provides a very suitable subset of multispectral bands for pixel classification purposes.

1. Introduction

Multispectral cameras record a specific portion of the electromagnetic spectrum in each band, therefore and by means of a narrow band-pass filter, each band captures a determined wavelength range producing a multi-band representation. The benefits of multispectral imaging in several disciplines is producing many emerging applications. Multi or hyperspectral sensors acquire information from a range of wavelengths in the spectrum and, unquestionably, they have produced an important improvement of the results obtained from just one or three bands in some demanding application fields, like remote sensing, medical imaging, product quality inspection, fine arts, etc.

Obviously, from the point of view of classification tasks, a very desirable step when we have a large amount of input spectral information is a process to reduce this initial information without losing classification accuracy in a significant way. This reduction could be done in two different ways: feature extraction [6, 4] or feature selection [1]. In feature extraction we would obtain a new and reduced data set representing the transformed initial information, whereas in feature selection we would have a subset of relevant data from the original information. In this work we will focus in feature selection rather than feature extraction due to the fact that in feature extraction the total amount of information is needed to obtain the new set of input bands. On the other hand, selecting the relevant range of wavelengths in the spectrum, where the process obtains better results, allows the acquisition step to deal with a reduced set and to make the analysis simpler.

In multispectral applications, the question is how select the correct bands from the multispectral range to characterise the problem. In this case, regarding to feature selection for pixel classification, this question could be addressed using information theory and, more concretely, by means of measures based on the mutual information concept [7].

The methodology of the algorithm presented in this work can be summarised as follows. A similarity space is defined among bands, where a dissimilarity measure is defined based on the mutual information between a pair of bands. From the initial set of bands that form a multispectral image, the process starts with a hierarchical clustering in the defined dissimilarity space, until reaching the K number of clusters of bands desired. In order to progressively construct a hierarchical family of derived clusters the method uses a linkage strategy with an inter-cluster distance as the objective function to optimise. Finally, a band representing each final cluster is chosen, which are considered the K most relevant bands.

2. Dimensionality reduction algorithm

In this section the dimensionality reduction algorithm is introduced. Let us remind that the objective is to reduce the initial number of bands, which represent the feature vector each pixel is represented, obtaining a smaller set of bands that provides as higher classification accuracy as possible, from the point of view of the pixel classification problem.

Therefore, the method proposed tries to identify the subset of bands that are as much independent as possible among them. It is known that independence between bands [6] is one of the key issues to obtain relevant subsets of bands for classification purposes. As we will show in the experimental results, trying to identify the subset of bands as much independent as possible among them, produces very satisfactory classification rates with respect other feature selection approaches.

[*] This work has been partly supported by projects ESP2005-07724-C05-05 from Spanish CICYT and P1-1B2004-36, P1-1B2004-08 from Fundació Caixa-Castelló

To find the subset of K selected bands that are as much independent as possible among them, the approach here presented defines a dissimilarity space based on mutual information between bands. In this dissimilarity space, a clustering process is performed. As a result of the clustering, bands are grouped according to the amount of information they share. Therefore, all the bands in the same cluster are highly dependent among them. In a final stage, a band representing each cluster is chosen, in such a way that the band selected will be the band that share as much information with respect to the other bands in the cluster. Eventually, the K selected bands from the final K clusters will have a significant degree of independence, and therefore, providing an adequate reduced representation that will provide satisfactory classification results.

2.1. Multispectral distance using mutual information

Let us calculate mutual information I from entropy measures according to the well-known expression $I(X,Y) = H(X)+H(Y)-H(X,Y)$, where $H(X)$, $H(Y)$ are the entropies of random vectors X, Y respectively and $H(X,Y)$ is the joint entropy. I is an absolute measure of common information between two sources, however, as we can infer from the previous equation, I by itself would not be a suitable distance measure. The reason is that it can be low because the X, Y variables present a weak relation (such as it should be desirable) or because the entropies are small (in such case, the variables contribute with few information). Thus, it is convenient to obtain a proper measure so that it works independently from the marginal entropies and also measures the statistical dependence as a distance.

Let us consider a set of n bands $X_1, ..., X_n$ from a multispectral image and let us suppose that each band represents a random variable. From this input data, we shall employ a measure of similarity between any two random images, $NI(X_i, X_j) = \frac{2 \cdot I(X_i, X_j)}{H(X_i)+H(X_j)}$, which is a normalised measure of I. This measure is used to calculate distance $D_{NI} = \left(1 - \sqrt{NI(X_i, X_j)}\right)^2$. Both D_{NI} and NI had been proposed in [2].

2.2. Hierarchical clustering

The hierarchical structures are a very intuitive way to summarise the input data. 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, that is, it starts with n initial clusters and, at each step, merges the two most similar groups to form a new cluster. Thus, the number of groups is reduced 1 by 1 until there are just the K clusters desired. To completely characterise our method we shall say that it is a hierarchical clustering algorithm based on a Ward's linkage method [8] and the distance D_{NI} described on section 2.1. Ward's linkage method has the property of producing minimum variance partitions. Thus, this method is also called minimum variance method because it pursues to form each possible group in a manner that would minimise the loss associated with each grouping (internal cohesion). To this end, the hierarchical grouping merges the pair of clusters that minimise the increment in the square error of the whole partition. The error used to this calculation is the intra-cluster dispersion. In addition to several studies that conclude that this method outperforms other hierarchical clustering methods [3], the process helps us to form groups with not much variance in their level of independence, that is, clusters with similar D_{NI} distances will be joined together.

Let us suppose that we merge clusters X_r and X_s. The general expression for the distance between the new cluster (X_r, X_s) and other cluster (X_k) is:

$$\begin{aligned} D[(X_k),(X_r,X_s)] &= \alpha \cdot D(X_k,X_r) + \beta \cdot D(X_k,X_s) + \\ &+ \gamma \cdot D(X_r,X_s) + \delta \cdot |D(X_k,X_r) - D(X_k,X_s)| \end{aligned} \tag{1}$$

where α, β, γ and δ are coefficients. Ward's inter-cluster distance results from the following coefficients, $\alpha = \frac{n_r+n_k}{n_r+n_s+n_k}$, $\beta = \frac{n_s+n_k}{n_r+n_s+n_k}$, $\gamma = \frac{-n_k}{n_r+n_s+n_k}$, $\delta = \emptyset$, where n_i is the number of instances in group i.

2.3. Choosing the cluster instances

The algorithm starts with the disjoint partition where each pattern (multispectral band) is a cluster. At this time, the distance matrix D is initialised by means of the distance explained on section 2.1. After that, the algorithm looks for the two most similar clusters that will have the minimum distance value in matrix D. Finally, these two clusters are merged into one and matrix D is updated using expression 1. Of course, the rows/columns corresponding to the merged clusters are deleted and a row/column for the new cluster is added.

The described 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. Thus, let us consider now the resulting cluster C with N bands. The weight of each band $i \in C$ is calculated as $W_i = \frac{1}{N} \sum_{j \in C, j \neq i} \frac{1}{\epsilon + D(i,j)^2}$, where ϵ is a very small value to avoid singular values, and function $D(i,j)$ returns the distance value between bands i,j. The representative band from each group is selected as the band with the highest W of the cluster. A low value of W_i means that the band i has an average large distance with respect to the other bands in the cluster, that is, in this case, the band i will have an average low correlation with respect 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 respect to the other bands in the cluster. Thus, choosing the band in the cluster with the highest average correlation, mutual information, with respect to the other bands in the cluster, what we are doing is choosing the band that better predicts the information content of the other bands, since the more mutual information two random variables have, the more can predict one of the variable about the other one.

As a result of the algorithm, there will be selected K bands that represent K clusters. These K bands will be significantly separated in the dissimilarity space defined, thus, having a low correlation (mutual information), and therefore having a high degree of independence among them.

3. Results

To test the proposed approach, different databases of multispectral images are used in the experimental results:

1. Multispectral images of oranges obtained by an imaging spectrograph (RetigaEx, Opto-knowledged Systems Inc., Canada). This database has two groups, VIS collection (400-720 nm in the visible) and NIR collection (650-1050 nm in the near infrared). In both cases, the camera has a spectral resolution of 10 nm. The database includes several kinds of orange defects. It has eight classes, obtaining 1463346 labelled pixels from VIS and 1491888 labelled pixels from NIR.

2. The $92AV3C$ source of data corresponds to a spectral image (145 X 145 pixels, 220 bands, 17 classes) acquired with the AVIRIS data set and collected in June 1992 over the Indian Pine Test site in Northwestern Indiana (http:/dynamo.ecn.purdue.edu /∼biehl/MultiSpec).

3. $DAISEX'99$ project provides useful aerial images about the study of the variability in the reflectance of different natural surfaces. This source of data corresponds to a spectral image (700 X 670 pixels, 6 classes) acquired with the 128-bands HyMap spectrometer during the DAISEX-99 campaign (http:/io.uv.es/projects/daisex/).

In addition to the previous description, images in Fig. 1 show some instances of the database collections used. These images are presented as RGB compositions.

In order to assess the performance of the method, a Nearest Neighbour (NN) classifier was used to classify pixels into the different classes. The performance of the NN classifier was considered as the validation criterion to compare the significance of the subsets of selected image bands obtained by the proposed approach and two supervised methods considered in the experiment carried out. Regarding to this last point, the main motivation about comparing the proposed method with supervised approaches is that the labelled data contains information about the distribution of classes existing in the hyperspectral data, and they allow the search for relevant feature subsets. Comparing the performance with those approaches, we can measure the capability to obtain subsets of relevant features (image bands) by the introduced algorithm without a prior knowledge of the class distributions in the multispectral image.

The first method is the well-known *ReliefF* algorithm [5] based on pattern distances. The second technique is related to divergence measures between classes. One of the best-known distance measures utilised for feature selection in multi-class problems is the average Jeffries-Matusita (JM) distance [1]. To obtain suboptimal subsets of features, we have applied a search strategy based on a Sequential Forward Selection applying this distance ($(SFS)JMdistance$).

In order to increase the statistical significance of the results, the experimental results shown in this section about the classification rates correspond to the average classification accuracy obtained by the NN classifier over five random partitions. The samples in each partition were randomly assigned to the training and test set with equal sizes as follows: VIS = 43902 pixels, NIR = 44758 pixels, HyMap = 37520 pixels, 92AV3C = 2102 pixels.

On the other hand, given the huge size of the data sets and the trouble in computational cost to apply the supervised approaches, particularly in the case of VIS, NIR and HyMap, the following independent partitions with respect to the data sets were randomly extracted maintaining the prior probability of the classes: VIS = 87805 pixels, NIR = 89516 pixels, HyMap = 93804 pixels and 92AV3C = 10512 pixels. Using these databases, the supervised approaches and the proposed method were applied in order to obtain a ranking of relevance of the features, that is, of bands.

Fig 2 represents the classification rate with respect to the subset of N bands selected by each method. In all cases, we show the performance of the *NN* classifier with respect to the number of features obtained by $WaLuMI$, $(SFS)JMdistance$ and $ReliefF$. Note that the proposed method obtained better performance with respect to the rest of methods in all databases. It is worthwhile mentioning that the $WaLuMI$ approach has a good behaviour in all cases when choosing the smaller sets of bands (one to ten), where the decision is more critical. Therefore, regarding to the band selection problem, where there exists high correlation among different features (image bands), the principle of looking for non-correlated bands from the different regions of the spectrum, by reducing the mutual information in the ensemble of image bands, has proven to be an effective approach to obtain subsets of bands that also provide results with satisfactory classification accuracy.

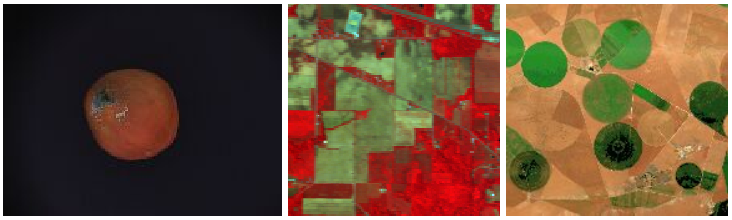

Figure 1. RGB composition examples. Orange image (VIS), 92AV3C image and HyMap image.

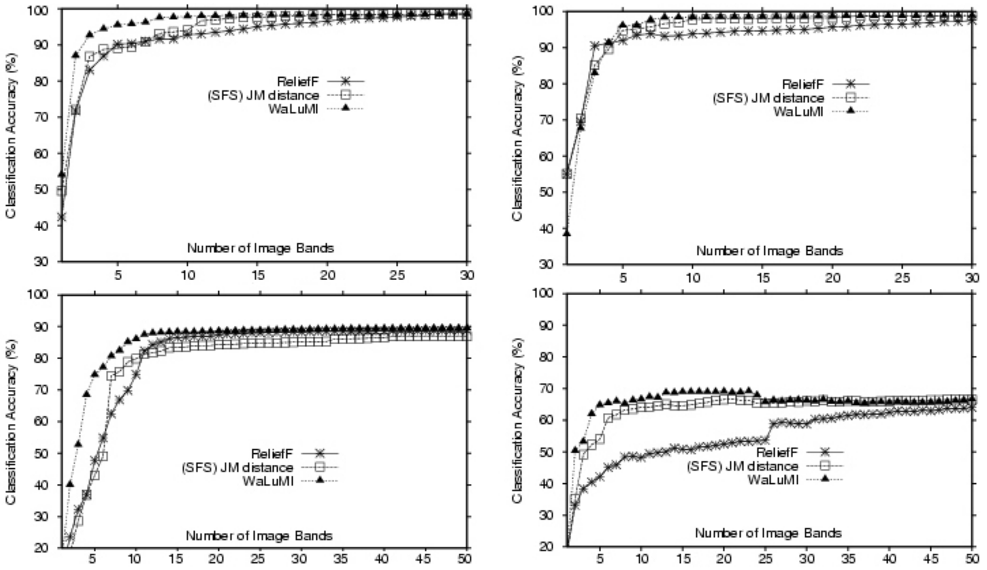

Figure 2. First row, VIS/NIR DB results. Second row, DAISEX'99 and 92AV3C DB results.

4. Conclusions

An approach to select bands in multispectral images based on mutual information has been introduced. It uses clustering to group bands correlated among them, selecting a subset of bands with a high degree of independence.

The results provide experimental evidence about the importance that independence among bands plays in the problem of classification. The method here presented is fully unsupervised and computationally affordable, avoiding the problem of labelling, and providing very satisfactory classification results with respect to other well-known supervised feature selection criteria.

References

[1] L. Bruzzonne, F. Roli, and S. S.B. An extension to multiclass cases of the jeffreys-matusita distance. *IEEE Transactions on Geoscience and Remote Sensing*, 33:1318–1321, 1995.

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

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

[4] L. Jimenez and D. Landgrebe. Supervised classification in high dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Transactions on System, Man, and Cybernetics*, 28, Part C(1):39–54, 1998.

[5] I. Kononenko. Estimating attributes: analysis and extensions of relief. *In Proceedings of 7th European Conference on Machine Learning, Catania, Italy*, pages 171–182, 1994.

[6] S. Kumar, J. Ghosh, and M. Crawford. Best basis feature extraction algorithms for classification of hyperspectral data. *IEEE Trans. on GRS*, 39(7):1368–1379, 2001.

[7] G. Tourssari, E. Frederick, M. Markey, and C. Floyd. Applications of mutual information criterion for feature selection in computer-aided diagnosis. *MLR*, 3:2394–2402, 2001.

[8] J. H. Ward. Hierarchical grouping to optimize an objective function. *Am Statistical Assoc*, 58(301):236–244, 1963.