Hyperspectral Data Selection from Mutual Information Between Image Bands

José Martínez Sotoca and Filiberto Pla

Dept. Llenguatges i Sistemes Informátics, Universitat Jaume I,
E-12071, Castelló, Spain
{sotoca, pla}@lsi.uji.es

Abstract. This work presents a band selection method for multi and hyperspectral images using correlation among bands based on mutual information measures. The relationship among bands are represented by means of the *transinformation matrix*. A process based on a Deterministic Annealing optimization is applied to the *transinformation matrix* in order to obtain a reduction of this matrix looking for the image bands as less uncorrelated as possible between them. Some experiments are presented to show the effectiveness of the bands selected from the point of view of pixel classification.

Keywords: Multispectral images, mutual information, deterministic annealing, unsupervised feature selection.

1 Introduction

Hyperspectral sensors acquire information in large quantities of spectral bands, which generate hyperspectral data in high dimensional spaces. These systems use multispectral image representations in order to estimate and analyze the presence of vegetation pathologies, substances or chemical compounds, pathologies, and so on, providing a qualitative and quantitative evaluation of those features.

When having available hyperspectral data, a common question to be solved is how to select the right spectral bands to characterize the problem. The main objective of band selection in multispectral imaging is to avoid redundant information and reduce the amount of data to be processed. Therefore, from the point of view of remote sensing, we would be interested in feature selection [3] rather than in feature extraction [7]. For instance, obtaining a new set of reduced image representations from a linear combination of the whole set of original image bands is not desirable, since we would need the total amount of information to obtain the new features. On the other hand, selecting a subset of relevant bands from the original set, allows the process of image acquisition to be reduced to a certain number of bands instead of dealing with the whole amount of data, making simpler the image acquisition and analysis.

In the framework of multispectral imaging, another possible answer to the problem of feature selection would be using an unsupervised approach [4][2]. In this work, a Deterministic Annealing (DA) approach is used to analyze the

D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109, pp. 853–861, 2006.

amount of information contained in the *mutual information matrix*, which represents the relations of information between pairs of spectral bands. The proposed algorithm uses a Deterministic Annealing (DA) approach to look for groups of bands as less correlated as possible, representing correlation between image bands by means of mutual information. Selected bands are further used in pixel classification tasks to assess the performance of proposed technique.

2 Deterministic Annealing for Rank Reduction

Let us consider a pair of random variables A_i and A_j, representing the image bands i and j. The amount of information contained in both images can be expressed as the joint entropy $H(A_i, A_j) = \sum p(a_i, a_j) \log_2 \frac{1}{p(a_i,a_j)}$, where $p(a_i, a_j)$ represents a joint probability distribution.

For two images i and j, the joint probability distribution $p(a_i, a_j)$ of both images can be estimated as $p(a_i, a_j) = \frac{h(a_i,a_j)}{MN}$, where $h(a_i, a_j)$ is the joint gray level histogram, and the normalizing factor, MN (M columns and N rows) is the image size.

Mutual information $H(A_i{:}A_j)$ is a basic concept in information theory [1]. It measures the interdependence between random variables. In the case of two images, the mutual information is defined as:

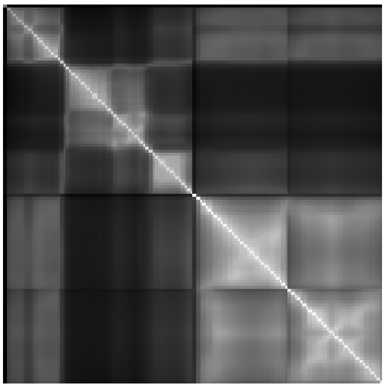
$$H(A_i : A_j) = H(A_i) + H(A_j) - H(A_i, A_j) \tag{1}$$

where $H(A_i)$, $H(A_j)$ are the entropy of images i and j, and $H(A_i, A_j)$ is the joint entropy.

One way to establish the interdependence between a set of features is defining the *transinformation matrix.* In our framework, this is a square matrix representing the mutual information between pairs of image bands. The diagonal terms represent the entropy of single bands, and contiguous bands in the spectrum tend to be highly correlated (brighter values in Fig 1).

The technique here proposed is aimed at reducing the rank of the *transinformation matrix* by selecting a given number of features that minimize the correlation among them. Therefore, we look for a global minimum without carrying out a search of subsets of features in the feature space. The process must be capable of picking up a subset of bands, in order to obtain as better performance as possible from the classification point of view reducing the feature space.

Discretizing the mutual information and representing the *transinformation matrix* as an "image" with gray levels (see Fig 1), defines a spreading measure of the information in the gray level distribution of the matrix. This measure will estimate the information contained about the appearance of the different regions of the spectrum in the matrix. Thus, we can analyze the probability that the event (value associated with each position of the matrix) takes place. This probability n_{ij} can be calculated as $n_{ij} = h_{ij}/D^2$, where h_{ij} is the value in the



(a)

Fig. 1. *Transinformation matrix* for a multispectral image with 128 wavebands. Darker values represent less correlation.

histogram for the gray level at i and j. Including the probability n_{ij} in each position in the matrix, we define the following function of information as:

$$I_{ij} = n_{ij} H(A_i : A_j). \tag{2}$$

From the function I_{ij} , we are interested in associating a probability of significance $p(I_{ij}|ij)$ for each position i and j in the matrix. This probability will mean how relevant is the interaction of band i and j for the problem. Therefore, a probabilistic model is applied over each position of the matrix $p(I_{ij}|ij)$. It is, thus, possible to utilize DA to obtain the image bands that contain higher values of significance in the matrix. To apply DA in such a framework, the following requirements must be fullfiled:

- The entropy S of the distribution of probabilities $p(I_{ij}|ij)$ associated to this representation of "level of uncertainty" must be maximum.
- The sum of probabilities are normalized to one.
- The product of $p(I_{ij}|ij)$ per the value of I_{ij} between pairs of bands, provides a value about the amount of information I associated to the ensemble.

Therefore, we can establish the the following relation:

$$S = -\sum_{i=1}^{D}\sum_{j=1}^{D} p(I_{ij}|ij) \log \frac{p(I_{ij}|ij)}{p_{ij}} \tag{3}$$

subject to

$$\sum_{i=1}^{D}\sum_{j=1}^{D} p(I_{ij}|ij) = 1 \quad \text{and} \quad \sum_{i=1}^{D}\sum_{j=1}^{D} p(I_{ij}|ij) I_{ij} = I \tag{4}$$

where p_{ij} is proportional to the prior contribution of each relation between pairs of bands. Thus, S is the entropy relative to some "measures" p_{ij} that has to be maximized [5]. To maximize S subject to the constraint Eq 4, we can introduce Lagrangian multipliers α and β,

$$S - \alpha \left(\sum_{i=1}^{D} \sum_{j=1}^{D} p(I_{ij}|ij) - 1 \right) - \beta \left(\sum_{i=1}^{D} \sum_{j=1}^{D} p(I_{ij}|ij) I_{ij} - I \right) \tag{5}$$

Setting the partial derivative of Eq 5 with respect $p(I_{ij}|ij)$ to zero, we obtain the following expression,

$$-\log p(I_{ij}|ij) - 1 + \log p_{ij} - \alpha - \beta I_{ij} = 0 \tag{6}$$

where

$$p(I_{ij}|ij) = p_{ij} e^{-\alpha - 1 - \beta I_{ij}} \tag{7}$$

Taking into account that the sum of probabilities are normalized to one, then

$$\sum_{i=1}^{D} \sum_{j=1}^{D} p_{ij} e^{-\beta I_{ij}} = e^{1+\alpha} = Z \tag{8}$$

where Z is the so-called *partition function* and

$$p(I_{ij}|ij) = \frac{p_{ij} e^{-\beta I_{ij}}}{Z} \tag{9}$$

Taking $\beta = \frac{1}{T}$, our probability function is expressed as

$$p(I_{ij}|ij) = \frac{p_{ij} e^{-I_{ij}/T}}{\sum_{i=1}^{D} \sum_{j=1}^{D} p_{ij} e^{-I_{ij}/T}}$$

and

$$p_{ij} = I_{ij} p(I_{ij}|ij)$$

The result is the Bayes' Theorem, where we can obtain the posterior probability distribution for each position through the exponential function of the values observed in the matrix multiplied by the prior probability p_{ij}.

The initialization of DA starts with large enough values of T, and a uniform distribution of probabilities $p(I_{ij}|ij) = 1/D^2$. The initial set of features X to choose is empty. As $T \rightarrow 0$ a reduction of the amount of information I is carried out. In practice, the system is annealed to a low temperature, such the amount of information I ("level of dependence" of the matrix) is sufficiently small.

On the other hand, we express the probability contributions of each band A_i accumulating for each row or column i (the matrix is symmetrical) as:

$$B_i = \sum_{j=1}^{D} p(I_{ij}|ij) \tag{10}$$

While T decreases, the difference between the values of $p(I_{ij}|ij)$ grow up. As T goes down, the probability contributions of some bands $B_i \rightarrow 0$, but it is possible that further in the annealing, with lower T, previous low values of B_i grow up for the new circumstances. Only if $B_i \cong 0$, we can almost assure that the corresponding band will not contribute in the probability distribution in the next iterations.

Summarizing, a brief sketch of the algorithm is as follows:

```
1. Initialize: $T = T_0$, $p(I_{ij}|ij) = 1/D^2$ and $|X| = 0$
2. Minimize: $F = I - TS$
3. Calculate: $B_i = \sum_{j=1}^{D} p(I_{ij}|ij)$
4. If $B_i \cong 0$ then: $X \leftarrow (X \cup A_i)$
5. Count the number of image bands $R$ such: $B_i > 1/D$
6. Lower Temperature: $T \leftarrow q(T)$
7. Go to step 2 while $R \geq 2$
```

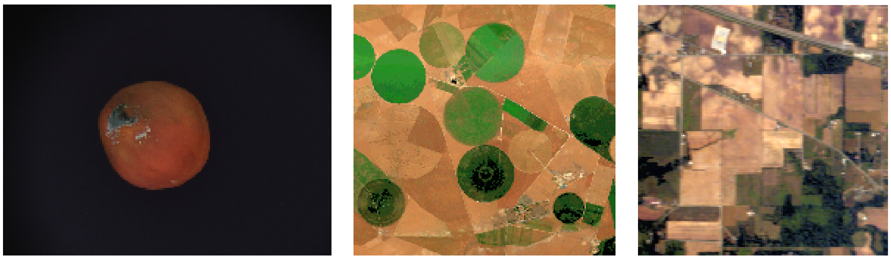
In our experiments, we used an exponential schedule to reduce T, $q(T) = \alpha T$, where $\alpha < 1$, but other annealing schedules are possible. At the end of the algorithm, the probability contributions B_i are concentrated in the two best bands with values about $\simeq 0.5$.

3 Empirical Results

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

1. Multispectral images of oranges obtained by an imaging spectrograph (Retiga-Ex, Opto-knowledge System Inc. Canada). This database was captured in to spectrum range, 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 nanometers. The database includes several kinds of defects. It has eight classes, obtaining 1463346 labelled pixels from VIS and 1491888 labelled pixels from NIR.
2. Spectral image (700 X 670 pixels) acquired with the 128-bands HyMap spectrometer during the DAISEX-99 campaign (http:/io.uv.es/projects/daisex/), and six different classes were considered in the area (see Fig 2 (b))
3. Spectral image (145 X 145 pixels) acquired with the AVIRIS data set with 220 bands collected in June 1992 over the Indian Pine Test site in Northwestern Indiana (see Fig 2 (c)). The data set is designated as *92AV3C*, and it has seventeen classes.(http:/dynamo.ecn.purdue.edu /∼biehl/MultiSpec)

In order to assess the performance of the method, a Nearest Neighbor (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 other methods (two supervised and one unsupervised approaches) in



(a) (b) (c)

Fig. 2. (a) Example of RGB composition for an orange image in the Visible spectrum. (b) HyMap RGB composition, Barrax, Spain. (c) RGB composition of AVIRIS (92AV3C: NW Indiana's Indian Pine test site).

the experiment carried out. To increase the statistical significance of the results, the average values over five random partitions were estimated.

In the case of supervised approaches, the main motivation is that the labelled data contains information about the distribution of classes exiting 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 DA approach without a prior knowledge of the class distributions in the multispectral image.

The first method is the well-known *ReliefF* algorithm [6] based on pattern distances. The second technique is related to divergence measures between classes. One of the best-known distance measures utilized for feature selection in multiclass problems is the average Jeffries-Matusita (JM) distance. To obtain suboptimal subsets of features, we have applied a search strategy based on a Sequential Forward Selection applying this distance ((SFS) JM distance) [3].

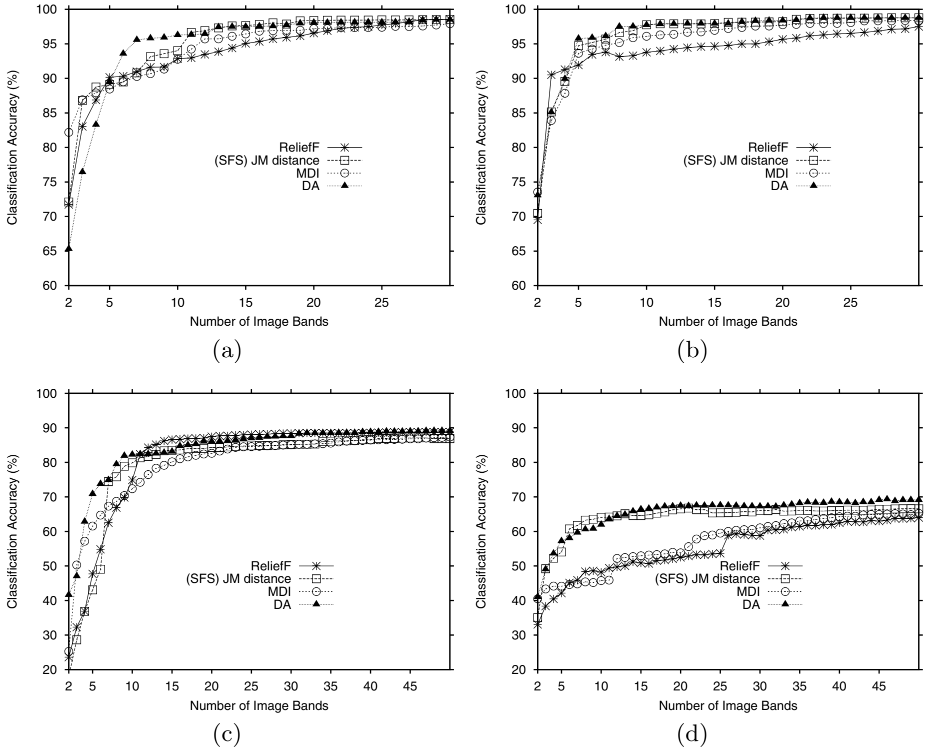
Moreover, we evaluated an unsupervised method presented in a previous work based in information measures between image bands [8]. This approach called "Minimization of the Dependent Information" (MDI) measures the region of dependence given a number bands for a multispectral image, and obtains a minimum interdependence.

3.1 Performance Evaluation

During the image labelling process, there are always pixels in an image that are not assigned to any class of interest, mainly because they are pixels that either do not clearly belong to some of the predefined classes or they are assigned to a complementary class. The pixels that have not been assigned to any class are labelled as "unknown" class.

The experimental results shown in this section about the classification rates correspond to the average classification accuracy obtained by the NN classifier over the five random partitions described previously. 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 proposed DA and the others methods were applied in order to obtain a ranking of relevance of the features, that is, of bands.



(a)

(b)

(c)

(d)

Fig. 3. (a) Results over oranges in VIS. (b) Results over oranges in NIR. (c) Results over spectral image with HyMap spectrometer. (d) Results over 92AV3C spectral image. In all cases, it is shown the performance of the *NN* classifier with respect to the number of features obtained by DA, (SFS) JM distance and *ReliefF*.

Fig 3 represents the classification rate with respect to the subset of N bands selected by each method. Note that the proposed DA method obtained better performance with respect to the rest of methods in the case of database of VIS, and similar accuracy for the best of the other approaches for NIR, HyMap and 92AV3C. It is worthwhile mentioning that the DA approach has a good behavior

Table 1. Computational cost in minutes (m) when selecting all features except for (SFS) JM distance, where it is shown for 30 features (VIS and NIR) and 50 features (HyMap and 92AV3C)

Criteria	Time (m)			
	VIS	NIR	HyMap	92AV3C
ReliefF	198 m	237 m	423 m	20 m
(SFS)JM distance	17 m	49 m	152 m	151 m
MDI	349 m	407 m	2337 m	2446m
DA	4 m	8 m	130 m	102 m

in all cases when choosing the smaller sets of bands (first one to ten), where the decision is more critical.

ReliefF performs poorer with respect to the other approaches except with HyMap image, where the performance of (SFS) JM distance is worse. *ReliefF* obtains a ranking of relevance for each single feature and the computational cost grows exponentially with respect to the number of samples in the data set.

(SFS) JM distance provides a high classification accuracy, but the computational cost grows exponentially with respect to the number of dimensions. Table 1 shows the computational time in minutes for the tested methods.

MDI provides similar classification accuracy respect to *ReliefF* but its nature is completely unsupervised. Moreover, it is not efficient from the computational point of view to obtain subsets in spaces with high dimensionality. This is mainly due to the cost of computing the joint probability distributions for each combination of bands.

In the case of DA, the principal problem arises when the *transinformation matrix* is built. Thus, the different co-occurrences of pixels in each pair of image bands are calculated [8], which represents an quadratic cost in time. On the other hand, when the matrix is built, the proposed DA method obtain the selected features very quickly.

Therefore, for 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 selected image bands that also provide satisfactory results from the classification accuracy point of view.

4 Concluding Remarks

In this work, correlation among image bands in multispectral images has been established in terms of mutual information. The relationships between bands can be represented by the *transinformation matrix*. Using this representation, an approach to rank reduction of the *transinformation matrix* using Deterministic Annealing has been proposed to look for a given number of bands as less correlated as possible among them.

Although the proposed method has not been established in terms of class separability for supervised training sets, it has been shown in the experimental results that the image bands selected by DA provide very satisfactory results with respect to classification accuracy when using the selected bands. This effect is more noticeable when choosing small sets of features, when the decision is more critical. These two advantages, its unsupervised nature and the ability to choose relevant bands in the case of small sets, represent the more relevant characteristics of the proposed approach.

Acknowledgements

This work has been supported in part by grants P1-1B2004-08 from Fundació Caixa Castelló-Bancaixa and ESP2005-07724-C05-05 from the Spanish CICYT.

References

1. J. Aczel, J., Daroczy, Z.: On measures of information and their characterization. New York: Academic Press, 1975.
2. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional for data mining applications. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, WA, June (1998), 94–105
3. Bruzzonne, L., Roli, F., Serpico S.B.: An extension to multiclass cases of the Jeffreys-Matusita distance. IEEE Transactions on Geoscience and Remote Sensing, **33** (1995) 1318–1321
4. Groves, P., Bajcsy, P.: Methodology for hyperspectral band and classification model selection. IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data. An Honorary Workshop for Prof. David A. Landgrebe, Washington D.C., 2003.
5. Jaynes, E.T.: Prior Probatilities. IEEE Transations on System Science and Cybernetic, SSC-4, (1968) pp. 227–241. Reprinted in Concepts and Applications of Modern Decision Models, V.M. Rao Tummala and R. C. Henshaw, eds., (Michigan State University Business Studies Series, 1976).
6. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In Proceedings of 7th European Conference on Machine Learning, Catania, Italy,(1994) 171–182
7. Kumar, S., Ghosh, J., Crawford, M.M.: Best basis feature extraction algorithms for classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, **39**, no. 7, (2001) 1368–1379
8. Sotoca, J.M., Pla F., Klaren A.C.: Unsupervised band selection for multispectral images using information theory. In 17th. International Conference on Pattern Recognition, Cambridge (UK),**3**, (2004) 510–513