Band selection using mutual information matrix for hyperspectral data

J.M. Sotoca, F. Pla
Dept. Llenguatges i Sistemes Informàtics, Universitat Jaume I
Av. Sos Baynat s/n, E-12071 Castelló de la Plana (Spain)
E-mail: {sotoca,pla}@uji.es

Abstract

In this paper, a band selection technique for hyperspectral image data is proposed. A mutual information matrix between pairs of bands is built to collect the relations of information between the different regions of the spectrum. A process based on a Deterministic Annealing optimization is applied on the mutual information matrix to obtain a probabilistic model and look for the image bands less uncorrelated as possible between them. Two supervised filter feature selection methods were also tested to analyze the accuracy obtained by the presented approach. The proposed methodology can develop for supervised selection, building the matrix in terms of class separability for labelled training sets.

1 Introduction

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

A multispectral image can be considered as defined in a $3D$ space $I(x, y, \lambda)$, where (x, y) denotes the spatial co-ordinates of the pixel location in the image, and λ denotes a spectral band (wavelength). Each spectral band records a specific portion of the electromagnetic spectrum so that each spectral band provides greater insight about the composition of the different regions of the image. Therefore, each image band is captured at the selected wavelength with a narrow band-pass filter, allowing a multi-band representation.

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 [8] rather than in feature extraction [10, 11]. 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. One way to solve it consists of grouping the data in the feature space by using clustering techniques [2]. Another approach is to minimize the classification error by selecting bands that provide the highest image contrast [5]. In this work, a Deterministic Annealing (DA) approach is used to analyze the amount of information contained in the *mutual information matrix*, which represents the relations of information for 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 Transinformation matrix

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 $H(A_i, A_j)$, that is,

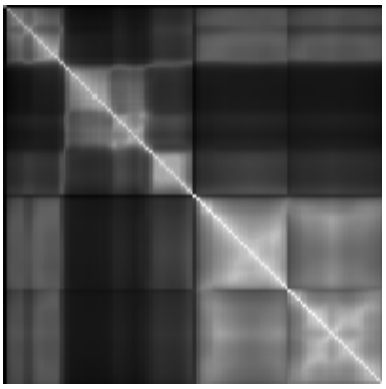
$$H(A_i, A_j) = \sum p(a_i, a_j) \log_2 \frac{1}{p(a_i, a_j)} \tag{1}$$

where $p(a_i, a_j)$ represents a joint probability distribution. The term $\log_2 \frac{1}{p(a_i,a_j)}$ means that the amount of information gained from a event with probability $p(a_i, a_j)$ is inversely related to the probability that this event take place. The rarer is an event, the more meaning is assigned to occurrence of the event. Thus, the information per event is weighted by the probability of occurrence. The resulting entropy term is the average amount of information gained from a set of possible events.

For two images i and j, the co-joint probability distribution $p(a_i, a_j)$ of both images can be estimated as,

$$p(a_i, a_j) = \frac{h(a_i, a_j)}{MN} \tag{2}$$

where $h(a_i, a_j)$ is the joint gray level histogram of both images, and the normalizing factor, MN (M columns and N rows) is the image size, assuming all images bands with equal size.



(a)

Figure 1: The Mutual Information matrix for a multispectral image with 128 wavebands. Darker values represent less correlation.

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:

$$H(A_i : A_j) = H(A_i) + H(A_j) - H(A_i, A_j) \tag{3}$$

where $H(A_i)$, $H(A_j)$ are the entropy of images i and j. The function $H(A_i{:}A_j)$ measures the amount of information shared between A_i and A_j. The entropies of both images satisfy the following inequality:

$$0 \leq H(A_i : A_j) \leq \min\{H(A_i), H(A_j)\} \tag{4}$$

One way to establish the interdependence between a set of features is defining the *transinformation matrix* (see Fig 1). This is a square matrix representing the mutual information between pairs of image bands. The diagonal terms represent the entropy of single band.

3 A new technique for rank reduction

We look for a strategy based on an unsupervised approach because, in supervised methods, it is necessary to fix beforehand the number of classes or regions present in the image, and to label the adequate number of training instances. Moreover, the computational cost of filter methods in supervised feature selection is considerable and, in many problems, labelling data can become a complex and difficult task.

Consider as input space the *transinformation matrix* with range D (number of spectral bands), representing the dependence among image bands. Contiguous bands in the spectrum tend to be highly correlated (brighter values in Fig 1). Looking at the *transinformation matrix*, we could interpret the problem of band selection as a rank reduction process of that matrix.

One possibility could be, for instance, to apply Truncate Singular Decomposition Value (TSVD) over the *transinformation matrix* or other factorization methods, eliminating the smaller singular values and their corresponding singular vectors. This idea has been used for noise reduction in signal processing [4].

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 few subset of bands in the mains regions that appear in the *transinformation matrix*, and obtaining as better performance as possible from the classification point of view reducing the feature space.

Given a certain function of information I_{ij} between pairs of bands represented in the matrix, 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. In the case of the *transinformation matrix*, each entry I_{ij} can represent the mutual information between bands.

On the other hand, discretizing I_{ij} values and representing them as gray levels (see Fig 1), allows to define a spreading measure of the information in the gray level distribution of the *transinformation matrix*. This measure will estimate the information contained about the appearance of the different regions of the spectrum in the matrix. Thus, we can consider the matrix as an "image" and analyze the probability that the event (value associate with each position of the matrix) take place. That is, the probability distribution associated to each position of the matrix n_{ij} can be calculated as $n_{ij} = h_{ij}/D^2$, where h_{ij} is the value in the histogram for the gray level at i and j.

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 uncertainly" 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:

$$S = -\sum_{i=1}^{D}\sum_{j=1}^{D} p(I_{ij}|ij)\log\frac{p(I_{ij}|ij)}{p_{ij}} \tag{5}$$

subject to

$$\sum_{i=1}^{D}\sum_{j=1}^{D} p(I_{ij}|ij) = 1 \quad \text{and} \quad \sum_{i=1}^{D}\sum_{j=1}^{D} p(I_{ij}|ij)I_{ij} = I \tag{6}$$

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 [6]. To maximize S subject to the constraint Eq 6, we can introduce Lagrangian multipliers α and β,

$$S + \alpha\sum_{i=1}^{D}\sum_{j=1}^{D} p(I_{ij}|ij) + \beta\sum_{i=1}^{D}\sum_{j=1}^{D} p(I_{ij}|ij)I_{ij} \tag{7}$$

Setting the partial derivative of Eq 7 with respect $p(I_{ij}|ij)$ to zero, we obtain the following expression,

$$-\log p(I_{ij}|ij) - 1 + \log p_{ij} + \alpha + \beta I_{ij} = 0 \tag{8}$$

where

$$p(I_{ij}|ij) = p_{ij}e^{\alpha-1+\beta I_{ij}} \tag{9}$$

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

$$\sum_{i=1}^{D}\sum_{j=1}^{D} p_{ij}e^{\beta I_{ij}} = e^{1-\alpha} = Z \tag{10}$$

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

$$p(I_{ij}|ij) = \frac{p_{ij}e^{\beta I_{ij}}}{Z} \tag{11}$$

On the other hand, we have to fix the Lagrangian multiplier β such as I and S are related. This optimization can be conveniently reformulated as the minimization of the following Lagrangian F with a parameter T:

$$F = I - TS \tag{12}$$

Therefore, the corresponding minimum of F is obtained by putting the Eq 11 into Eq 12

$$F^* = \min(F) = -T\log\left(\sum_{i=1}^{D}\sum_{j=1}^{D} p_{ij}e^{\beta I_{ij}}\right) \tag{13}$$

Multiplying $p(I_{ij}|ij)$ in Eq 8 and adding for all values, we obtain

$$-\sum_{i=1}^{D}\sum_{j=1}^{D} p(I_{ij}|ij)\log\frac{p(I_{ij}|ij)}{p_{ij}} - (1-\alpha)\sum_{i=1}^{D}\sum_{j=1}^{D} p(I_{ij}|ij) + \beta\sum_{i=1}^{D}\sum_{j=1}^{D} p(I_{ij}|ij)I_{ij} = 0 \tag{14}$$

then

$$S + \beta I = 1 - \alpha = lnZ \tag{15}$$

and from the Eq 12

$$S - \frac{I}{T} = -\frac{F}{T} \tag{16}$$

Thus, we can consider $\beta = -1/T$ and $lnZ = -F/T$. Finally, 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

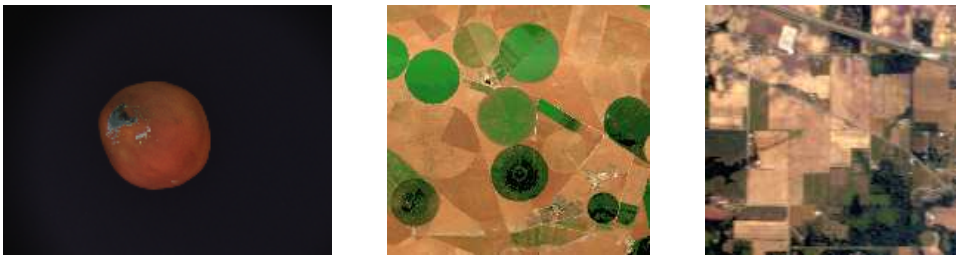
$$p_{ij} = I_{ij}p(I_{ij}|ij)$$

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 per the prior probability p_{ij}.

In our experiments, we have observed that using $I_{ij} = H(A_i{:}A_j)$, the approach finds a global minimum in regions of the spectrum of image bands with smaller values of mutual information with respect to the rest of regions of the spectrum represented in the *transinformation matrix*. Nevertheless, to obtain a good performance of the classifier in the subset of features selected, it is necessary to choose image bands of the different regions more representative of the ensemble. This question can be solved introducing the probability to appear this event in the matrix n_{ij} in the function of information as:

$$I_{ij} = n_{ij}H(A_i : A_j). \tag{17}$$

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. It is clear from Eq 12 that the goal at each temperature is to maximize the entropy of the partition. As $T \to 0$ 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.



(a) (b) (c)

Figure 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).

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:

$$B_i = \sum_{j=1}^{D} p(I_{ij}|ij) \tag{18}$$

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 and 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 $\simeq 0.5$.

4 Experiments and results

In the experiments, we have used four sources of hyperspectral or multispectral data. The two first collections of multispectral images were obtained by an imaging spectrograph (Retiga-Ex, Opto-knowledge System Inc. Canada). The first one has a spectral range from 400 to 720 nanometers in the visible (VIS) obtaining a set of 33 spectral bands for each image. The second one has a spectral range from 650 to 1050 nanometers in the near infrared (NIR) obtaining a set of 41 spectral bands for each image. In both cases, the camera has a spectral resolution of 10 nanometers.

The image database consisted of forty multispectral images for the VIS and NIR, respectively, corresponding to orange fruits with different types of defects and skin variations on their surfaces (see Fig 2 (a)). In order to compare the performance of the approach here presented, different region of the oranges, including the background, were labelled in eight classes, obtaining 1463346 pixels from VIS and 1491888 pixels from NIR.

The third source of data corresponds to a spectral image (700 X 670 pixels) acquired with the 128-bands HyMap spectrometer during the DAISEX-99 campaign with six different classes were considered in the area (see Fig 2(b)) (http:/io.uv.es/projects/daisex/).

The fourth source of data corresponds to a 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 two supervised methods considered in the experiment carried out. To increase the statistical significance of the results, the average values over five random partitions were estimated.

4.1 Supervised criteria proposed

To analyze the accuracy of the ranking of bands obtained by the proposed approach, two supervised filter feature selection methods were also tested. Thus, the band selection process was considered as a supervised feature selection approach, in this case using the labelled data set for the feature selection process.

The main motivation about comparing the proposed method with supervised approaches 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 [9] based on pattern distances. This algorithm initializes every feature weight to zero and then iterates m times looking for a set of feature weights that optimizes a criterion function.

The procedure begins by randomly selecting a sample x from the data set. For the selected sample, it determines the nearest neighbor prototype of the same class p^{hit} (nearest hit) and the nearest neighbor prototype of the different class p^{miss} (nearest miss). The algorithm updates each feature weight f_i according to the following criterion:

$$f_i^m = f_i^{m-1} - \frac{diff(x_i, p_i^{hit})}{m} + \sum_{c \neq class(x)} \frac{p(c) diff(x_i, p_i^{miss})}{m} \tag{19}$$

where $p(c)$ is the prior probability of class c, and $diff(,)$ is the distance between the sample and the prototype for the feature i. This algorithm was chosen because of its widespread use and good performance in general feature selection problems. As a result, the higher weight, the more relevant is a feature.

The second technique is related to divergence measures between classes. One of the best-known distance measures utilized for feature selection in multi-class problems is the average Jeffries-Matusita (JM) distance [8]:

$$JM = \sum_{h=1}^{c} \sum_{k>h}^{c} P_h P_k JM_{hk} \tag{20}$$

where

$$JM_{hk} = \sqrt{2(1 - e^{-b_{hk}})}$$

and

$$b_{hk} = \frac{1}{8}(m_h - m_k)^t \left(\frac{S_W^h + S_W^k}{2} \right)^{-1} (m_h - m_k) + \frac{1}{2} \log \left(\frac{\left| \frac{S_W^h + S_W^k}{2} \right|}{\sqrt{|S_W^h||S_W^k|}} \right)$$

P_i is the priori probability of the i-th class, b_{hk} is the Bhattacharyya distance between the classes h and k. S_W^i and m_i are the covariance matrix and the mean vector of the class i, respectively.

In terms of class separability, the higher is the JM distance between two classes, the more separability between them. To obtain suboptimal subsets of features, we have applied a search strategy based on a Sequential Forward Selection applying this distance ((SFS) JM distance). This technique starts from an empty feature subset and adding one feature at a time, reaching a feature subset with the desired cardinality.

4.2 Experiments including background pixels

During the image labelling process, there is 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 "background" class. In this subsection, we include the background information in the databases for its evaluation.

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 supervised approaches and the proposed DA method were applied in order to obtain a ranking of relevance of the features, that is, of bands.

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 other three databases (NIR, HyMap and 92AV3C). It is worthwhile mentioning that the DA approach has a good behavior 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.

On the other hand, (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.

In the case of DA, the principal problem arises when we build the *transinformation matrix*. Thus, the different co-occurrences of pixels in each pair of image bands are calculated [7], which represents an important 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

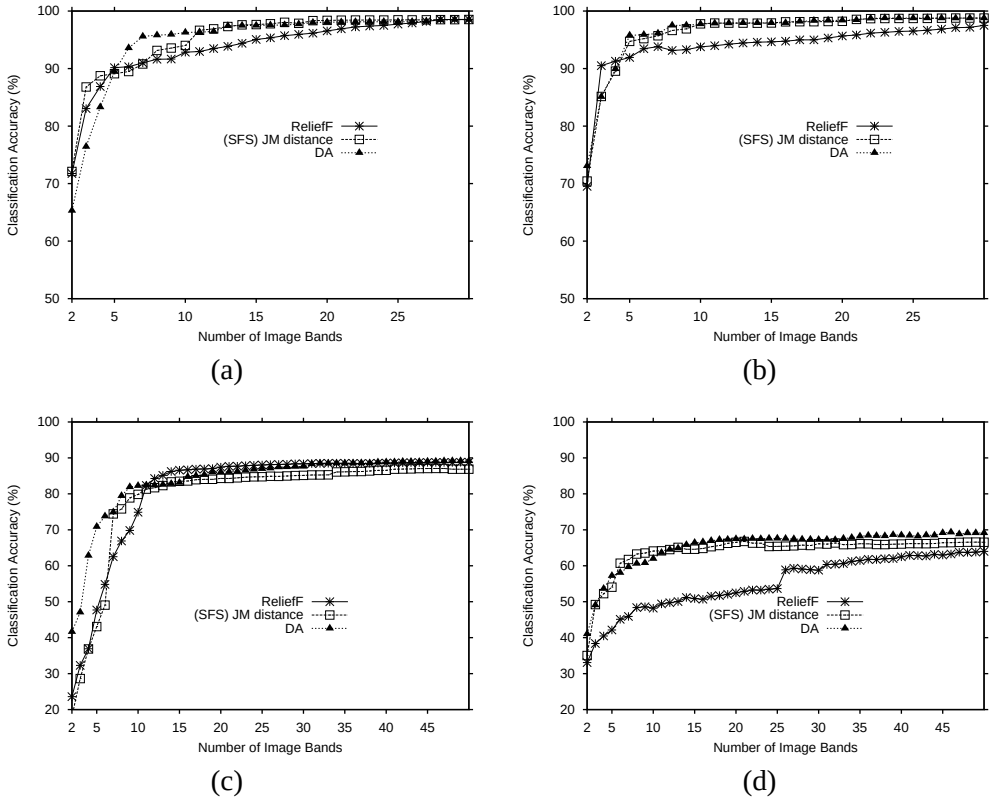


Figure 3: (a) Results over oranges in VIS. (b) Results over oranges in NIR. (c) Results over spectral image with HyMap spectrometer. (d) Results over 92AV3C spectral image. In all cases, it is shown the performance of the *NN* classifier with respect to the number of features obtained by DA, (SFS) JM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

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
DA (build the matrix)	4 m	8 m	130 m	102 m

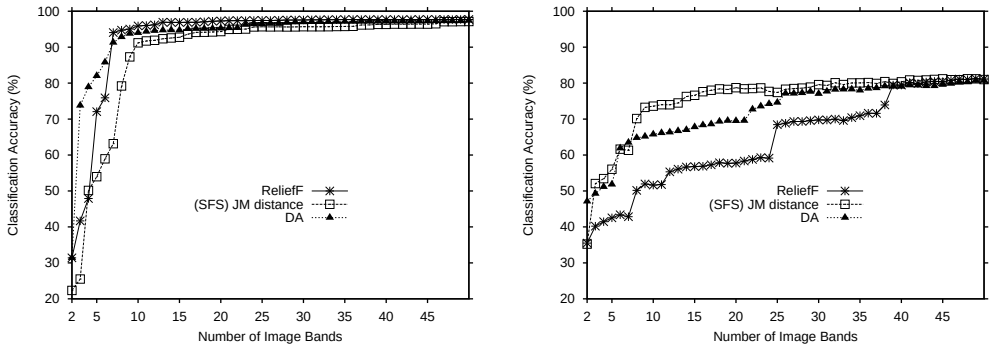


Figure 4: (a) Results over spectral image with HyMap spectrometer. (b) Results over 92AV3C spectral image. In all cases, we show the performance of the *NN* classifier with respect to the number of features obtained by DA, (SFS) JM distance and $ReliefF$.

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.

4.3 Experiments without background pixels

The hyperspectral data assigned to the "background class" are usually very scattered and overlapped with other classes, and this fact damages the classification accuracy. Moreover, the elimination of this information supposes a supervised knowledge to detect those regions of the image.

These regions are very difficult to detect with precision from unsupervised information. Therefore, the goal of this experiment is analyzing the advantages that suppose the knowledge of the class distribution without the noise that the *background* class can introduce. In this case, we will focus on HyMap and 92AV3C hyperspectral data, where the background information is much more undefined.

In the case of HyMap, we added the *background* class to the training set and validation set: training = 26190 pixels and validation = 65479 pixels. The test set contains all classes except the *background* class. The total number of test samples is 327336 pixels. Thus, the experiment classifies the test using the ranking of relevance of the features obtained by the validation set with the proposed method and the supervised methods used in the comparison.

The image 92AV3C only contains 10366 instances without the *background* class. Therefore, we apply a holdout partition, where the training and the validation set have the same size with 5181 pixels and the rest of pixels represent the test set = 5185 pixels.

In Fig 3 (a), the best performance is obtained by $Relief$ over HyMap, although DA

reaches a good performance, even better in the first two to seven bands, where the decision is more critical. (SFS) JM distance provides the worst accuracy of the three methods. On the other hand, $Relief$ needs 13 features to reach 96.94 % , while similar experiments realized by Camps et. al. [3] using Support Vector Machines (SVM) only needs 2 features reaching 96.44 %. In this sense, the NN classifier degrades more rapidly than SVM, when the dimension of the input space is lower.

In the case of the image 92AV3C, the NN classifier achieves the best performance using the ranking obtained by (SFS) JM distance. In this case, it exits a clear improving for this method with respect to the other ones. Therefore, the knowledge of the spatial distributions of the sixteen classes allows a better search to pick up goods subset of features.

5 Conclusions and future research

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 highly relevant bands in the case of small sets, represent the more relevant characteristics of the proposed approach.

Acknowledgments

This has work been supported in part by grants IST-2001-37306 (IST Project European Union) and P1-1B2004-08 from Fundació Caixa Castelló-Bancaixa.

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] Camps-Valls, G., Gómez-Chova, L., Calpe-Maravilla, J., Soria-Olivas, E., Martín-Guerrero, J.D., Moreno J.: Support Vector Machines for Crop Classification Using Hyperspectral Data. In 1st. Iberian Conference on Pattern Recognition and Image Analysis, Mallorca, Spain, (2003) 134-141

[4] Hansen, P.C., and Jensen, S.H.: FIR Filter Representation of Reduced-Rank noise Reduction. IEEE Transaction On Signal Processing, **46** (1998) 1737–1741

[5] 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.

[6] 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).

[7] 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

[8] 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

[9] Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In Proceedings of 7th European Conference on Machine Learning, Catania, Italy,(1994) 171–182

[10] 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

[11] Jimenez, L., Landgrebe, D.: Supervised classification in high dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data. IEEE Transactions on System, Man, and Cybernetics, **28**, Part C., no. 1, (1998) 39–54