INT. J. REMOTE SENSING, 2000, VOL. 21, NO. 3, 549–563

A technique for feature selection in multiclass problems

L. BRUZZONE and S. B. SERPICO

Department of Biophysical and Electronic Engineering, University of Genoa, Via Opera Pia, 11a, I-16145 Genova, Italy; e-mail address: lore@dibe.unige.it

(*Received 29 April 1998; in final form 22 December 1998*)

Abstract. One of the main phases in the development of a system for the classification of remote sensing images is the definition of an effective set of features to be given as input to the classifier. In particular, it is often useful to reduce the number of features available, while saving the possibility to discriminate among the different land-cover classes to be recognized. This paper addresses this topic with reference to applications that involve more than two land-cover classes (multiclass problems). Several criteria proposed in the remote sensing literature are considered and compared with one another and with the criterion presented by the authors. Such a criterion, unlike those usually adopted for multiclass problems, is related to an upper bound to the error probability of the Bayes classifier. As the objective of feature selection is generally to identify a reduced set of features that minimize the errors of the classifier, the aforementioned property is very important because it allows one to select features by taking into account their effects on classification errors. Experiments on two remote sensing datasets are described and discussed. These experiments confirm the effectiveness of the proposed criterion, which performs slightly better than all the others considered in the paper. In addition, the results obtained provide useful information about the behaviour of different classical criteria when applied in multiclass cases.

1. Introduction

The availability of automatic classification systems devoted to producing reliable and accurate thematic maps is an important requirement for the development of numerous remote sensing applications. To this end, it is fundamental to provide a classifier with effective features to distinguish accurately among different land-cover classes. Such features may be obtained both by developing new types of sensors (e.g. hyperspectral sensors (Richards 1993, Schowengerdt 1997)) and by improving the techniques for extracting information from images acquired in different spectral channels (e.g. by computing texture features, vegetation indexes, etc. (Richards 1993, Schowengerdt 1997)). In both cases, a large amount of features can be obtained that are usually strongly correlated, particularly when one uses sensors with high spectral resolution that acquire images in a large number of very close spectral bands. The consequent redundancy of the feature set (i.e. the presence of features containing similar information) suggests reducing the number of features given as input to a classifier, while maintaining classification accuracy as high as possible (Swain and Davis 1978, Fukunaga 1990, Richards 1993). A reduction in the number of features given as input to a classifier makes it possible to decrease both the computational

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online
http://www.tandf.co.uk/journals/tf/01431161.html

time required by the classification process and, in some cases, the costs of the computation of features and of the storage of the images from which features are extracted. Moreover, a reduction in the number of features may also increase classification accuracy (Hughes phenomenon) (Fukunaga 1990).

Several techniques aimed at reducing the number of features have been proposed in the literature (Kailath 1967, Swain and Davis 1978, Thomas *et al.* 1987, Fukunaga 1990, Mausel *et al.* 1990, Richards 1993). In particular, two different approaches can be defined:

(i) It is possible to extract some features that compress the information contained in the original features through the application of appropriate linear or nonlinear transformations to the original feature space. This approach is usually named *feature extraction*;
(ii) It is possible to derive a subset of the original set of features that allow one to separate accurately the land-cover classes considered. This approach is usually called *feature selection*.

The techniques belonging to the first approach (e.g. *Principal Component Analysis* (Richards 1993, Schowengerdt 1997) and the *Decision Boundary* method (Lee and Landgrebe 1993)) have the advantage of compressing the information available in the original feature set into a subset of uncorrelated features. Unfortunately, they exhibit the drawback of losing the physical significance of features. At the end of the transformation process, it is difficult to understand which real physical parameters are used by the classifier to distinguish the information classes.† In some remote sensing applications, this fact may represent a limitation on the understanding of the behaviour of the implemented classification system and hence on the validation of its performance.

The second approach generally involves both a search algorithm and a criterion function. The search algorithm generates and compares possible 'solutions' of the feature-selection problem (i.e. subsets of features) by utilizing the criterion function as a measure of the effectiveness of each considered feature subset. The best feature subset found in this way is the output of the feature-selection algorithm.

As the objective of feature selection is to select features that minimize the overall error of the classifier, it would be appropriate to adopt criterion functions related to the behaviour of the error made by the classifier used. But, in multiclass remote sensing problems, only in a few cases do the adopted criterion functions exhibit such a characteristic (Bruzzone *et al.* 1995). This is true, even though, within the framework of information theory, several works defined bounds to the error probability in multiclass cases (Lainiotis 1969, Hellman and Raviv 1970, Lainiotis and Park 1971, Devijver 1974, Garber and Djouadi 1988). This depends on the fact that most of such bounds are rather complex; therefore, using them in real applications is not practical (Devijver 1974). In this sense, it seems interesting to consider criteria that are related to the error probability, and that are simple enough from a computational viewpoint.

Concerning the search algorithm (Jain and Zongker 1997), both optimal strategies (e.g. the *branch and bound* technique (Fukunaga 1990)) and suboptimal strategies

†It is worth noting that it is usually possible to associate particular properties of land covers with the response in each spectral band.

(e.g. the *sequential forward selection* and the *sequential backward selection* (Pudil *et al.* 1994), *floating search* methods (Pudil *et al.* 1994), and *genetic algorithms* (Siedlecki and Slansky 1989)) have been proposed. After fixing the desired number of features in the selected set, optimal strategies identify the best set of features according to the adopted criterion function. Unfortunately, these strategies often turn out to be unsuitable from a computational viewpoint, due to their intrinsic combinatorial complexity (Jain and Zongker 1997). In such cases, suboptimal strategies can be adopted, which allow a good set of features (but not necessarily the best one) to be selected.

In this paper, we focus attention on criterion functions devoted to feature selection in multiclass cases for classification of remote sensing images acquired by passive sensors. The choice of considering multiclass cases stems from the fact that they are very frequent in remote sensing problems; the interest in passive sensors results from the wide use of such sensors in real applications. In particular, we present a simple, yet effective, criterion function that can be adopted in multiclass cases, and that is related to an upper bound to the error probability of the Bayes classifier, under the hypothesis of Gaussian distribution (commonly assumed for passive-sensor data). In addition, we report experimental results obtained on two remote sensing datasets to compare the performance of the above-mentioned criterion function with those of several feature-selection techniques commonly used in remote sensing for multiclass problems.

2. Feature selection in multiclass problems

For feature selection in multiclass problems, one can adopt criterion functions that have been defined for multiclass cases, or one can consider criterion functions originally defined for two-class cases and generalize them to multiclass ones. In the following, attention will be focused on some criterion functions that are among the most widely used in remote sensing applications.

Criterion functions are typically measures of the statistical separability of classes in a given feature space. For remote sensing applications, the divergence criterion (Swain and Davis 1978, Thomas *et al.* 1987, Mausel *et al.* 1990, Richards 1993), the transformed divergence criterion (Swain and Davis 1978, Thomas *et al.* 1987, Mausel *et al.* 1990, Richards 1993), the Bhattacharyya distance (Kailath 1967, Swain and Davis 1978, Thomas *et al.* 1987, Mausel *et al.* 1990, Richards 1993), the Jeffreys–Matusita (J–M) distance (Swain and Davis 1978, Thomas *et al.* 1987), and the criteria based on scatter matrices (Fukunaga 1990, Liu and Jernigan 1990) are the most widely used criteria. Our analysis will be restricted to the separability indexes based on the Bhattacharyya distance, the J–M distance, and one from among a number of possible separability measures based on scatter matrices because these separability indexes are representative of the main types of feature-selection criteria.

Let us consider a classification problem in which each pattern, described by an n-dimensional feature vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ in the feature space X, is to be assigned to one of c different classes $\Omega = (\omega_1, \omega_2, \ldots, \omega_c)$ characterized by the *a priori* probabilities $P(\omega_i)$ $(i = 1, 2, \ldots, c)$. Let $p(\boldsymbol{x}/\omega_i)$ be the conditional probability density functions for the feature vector $\boldsymbol{x}$, given the class ω_i $(i = 1, 2, \ldots, c)$. The objective of feature selection is to select the best subset of m features (with $m < n$) according to the optimization of the criterion function (e.g. maximization of a separability index or minimization of an error bound).

Various authors have proposed both theoretically based and empirically

developed generalizations of two-class distances to multiclass cases. The most common strategy is to use the weighted average distances computed for all pairs of classes (Swain and Davis 1978, Thomas *et al.* 1987, Mausel *et al.* 1990, Richards 1993).

Applying this strategy to the Bhattacharyya distance (Kailath 1967), we obtain:

$$B_{ave} = \sum_{i=1}^{c} \sum_{j=1}^{c} P(\omega_i)P(\omega_j)B_{ij} \tag{1}$$

where B_{ij} is the Bhattacharyya distance between two classes, ω_i and ω_j, and can be expressed as (Kailath 1967, Swain and Davis 1978, Fukunaga 1990, Mausel *et al.* 1990):

$$B_{ij} = -\ln\left\{\int_x \sqrt{p(\boldsymbol{x}/\omega_i)p(\boldsymbol{x}/\omega_j)}\, d\boldsymbol{x}\right\} \tag{2}$$

B_{ij} represents a measure of the average statistical distance between the conditional probability density functions related to the two classes. For remote sensing data acquired by passive sensors, it is usually assumed that classes have Gaussian distributions. In this case, B_{ij} can be simplified (Swain and Davis 1978) as:

$$B_{ij} = \frac{1}{8}(\boldsymbol{m}_i - \boldsymbol{m}_j)^t \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\boldsymbol{m}_i - \boldsymbol{m}_j) + \frac{1}{2}\log\left[\frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{\sqrt{|\Sigma_i||\Sigma_j|}}\right] \tag{3}$$

where $\boldsymbol{m}_i$, $\boldsymbol{m}_j$ and Σ_i, Σ_j are the mean vectors and the covariance matrices, respectively, for the classes ω_i and ω_j.

The same generalization strategy can be applied to the J–M distance:

$$J_{ave} = \sum_{i=1}^{c} \sum_{j=1}^{c} P(\omega_i)P(\omega_j)J_{ij} \tag{4}$$

where J_{ij}, that is, the J–M distance between ω_i and ω_j, is defined as (Swain and Davis 1978, Thomas *et al.* 1987):

$$J_{ij} = \left\{\int_x [\sqrt{p(\boldsymbol{x}/\omega_i)} - \sqrt{p(\boldsymbol{x}/\omega_j)}]^2\, d\boldsymbol{x}\right\}^{1/2} \tag{5}$$

Under the assumption of Gaussian distribution of classes, also the above distance can be simplified, as the J–M distance J_{ij} between two classes can be rewritten as a function of the Bhattacharyya distance B_{ij} between the two classes (Swain and Davis 1978, Thomas *et al.* 1987, Mausel *et al.* 1990):

$$J_{ij} = [2(1 - e^{-B_{ij}})]^{1/2} \tag{6}$$

and B_{ij} can be computed according to equation (3).

One can perform feature selection by selecting the feature subset that maximizes equations (1) or (4). Unlike what happens in two-class cases, in multiclass cases the Bhattacharyya and the J–M criteria may select different subsets of features (Swain and Davis 1978). This depends on the fact that the Bhattacharyya distance between two classes continues to increase significantly even when the topological distance between them (e.g. the distance between the mean vectors) reaches values corresponding to well-separated classes. On the contrary, the J–M distance, by analogy to the error probability, exhibits a 'saturating' behaviour for large distance values. Thanks

to this similarity of the J–M distance to the error probability behaviour, the criterion based on the J–M distance is usually more effective (Swain and Davis 1978). Accordingly, for the other generalizations to multiclass cases considered in the following, we shall focus attention just on the J–M distance.

In a previous paper (Bruzzone *et al.* 1995), the authors pointed out that the J–M distance can be applied to multiclass cases, according to the Bhattacharyya bound to the Bayes error, as follows:

$$J_{bh} = \sum_{i=1}^{c} \sum_{j>i}^{c} \sqrt{P(\omega_i)P(\omega_j)} J_{ij}^2 \tag{7}$$

One can use the J–M distance for feature selection in multiclass cases also by selecting the set of features that maximize the separability index J_{min} given by (Swain and Davis 1978):

$$J_{min} = \min_{i,j} \{J_{ij}\} \qquad i = 1,\ldots, c;\ j = 1, \ldots, c;\ i \neq j \tag{8}$$

that is, the set of features that can best separate the least distinguishable pair of classes are selected. The analysis of J_{ave}, J_{bh} and J_{min} highlights how the different generalizations of the J–M distance to multiclass cases may provide different results. J_{ave} is the most classical generalization typically used in the literature. J_{bh}, being based on the Bhattacharyya upper bound to the Bayes error probability, is the only generalization that makes it possible to maintain a theoretical relation between selected features and error probability. Moreover, as discussed in Bruzzone *et al.* (1995), such generalization gives greater importance to data classes with low *a priori* probabilities in the selection process, as compared with J_{ave}. On the contrary, J_{min}, which is based on an entirely different concept, guarantees the accurate separation of the two most critical classes, but it may consequently lead to the selection of features that are not effective for other classes.

Other feature-selection criteria evaluate the effectiveness of features by computing within-class (S_w) and between-class (S_b) scatter matrices, defined in a general multiclass case as (Fukunaga 1990):

$$S_w = \sum_{i=1}^{c} P(\omega_i)\Sigma_i \tag{9}$$

$$S_b = \sum_{i=1}^{c} P(\omega_i)(\boldsymbol{m}_i - \boldsymbol{m}_0)(\boldsymbol{m}_i - \boldsymbol{m}_0)^t \tag{10}$$

where c is the number of classes and $\boldsymbol{m}_0$ denotes the expected vectors of the 'mixture' distribution given by:

$$\boldsymbol{m}_0 = \sum_{i=1}^{c} P(\omega_i)\boldsymbol{m}_i \tag{11}$$

From the matrices S_w and S_b, several separability indexes can be derived (Fukunaga 1990, Liu and Jernigan 1990). As an example, a separability index is the following (Liu and Jernigan 1990):

$$F = \frac{|S_w + S_b|}{|S_w|} \tag{12}$$

This index evaluates the effectiveness of features by considering their capability to provide a large inter-class separation and a small intra-class spread by analysing

together samples of all classes. Even though the indexes based on scatter matrices are widely used, they present the drawback of not exhibiting a saturating effect for large distance values.

When data acquired by passive sensors are used, all the aforementioned criteria can be adopted, under the reasonable assumption of Gaussian distributions of classes. Only the criterion based on scatter matrices does not explicitly require any hypothesis on class distributions; however, it is effective only if classes have unimodal and symmetric distributions (Fukunaga 1990). It is worth noting that when no simple assumptions on the distributions of classes can be made, it is necessary to apply criteria suited to solving problems characterized by non-parametric or multimodal distributions (Novovicova *et al.* 1996, Krishnan *et al.* 1996).

3. The proposed criterion

The criterion we present in this paper is based on an upper bound to the Bayes error formulated under appropriate simplifying hypotheses. We define the criterion for two-class cases and then generalize it to multiclass cases.

Let us consider two classes, ω_i and ω_j. The error probability of the Bayes classifier for the minimum error is given by (Tou and Gonzalez 1974, Fukunaga 1990):

$$P_e(\omega_i, \omega_j) = P(\omega_i) \int_{x \in D_j} P(\xi/\omega_i) d\xi + P(\omega_j) \int_{x \in D_i} P(\xi/\omega_j) d\xi \tag{13}$$

where D_i and D_j are the 'decision regions' in the feature space X for the classes ω_i and ω_j, respectively, and are defined as:

$$D_i = \{\boldsymbol{x} \in X \,|\, P(\omega_i) p(\boldsymbol{x}/\omega_i) \geqslant P(\omega_j) p(\boldsymbol{x}/\omega_j) \tag{14}$$

$$D_j = \{\boldsymbol{x} \in X \,|\, P(\omega_j) p(\boldsymbol{x}/\omega_j) > P(\omega_i) p(\boldsymbol{x}/\omega_i)\} \tag{15}$$

Under the hypotheses of Gaussian distributions and of two classes with equal covariance matrices (i.e. $\Sigma_i = \Sigma_j = \Sigma_{ij}$), equation (13) can be rewritten as (Tou and Gonzalez 1974):

$$P_e(\omega_i, \omega_j) = P(\omega_i) \left[1 - Q\left(\frac{\alpha - \frac{1}{2} d_{ij}}{\sqrt{d_{ij}}} \right) \right] + P(\omega_j) Q\left(\frac{\alpha + \frac{1}{2} d_{ij}}{\sqrt{d_{ij}}} \right) \tag{16}$$

where $Q(x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_x^{+\infty} e^{-\xi^2/2} d\xi$, the value α depends on the optimal decision threshold computed using the maximum *a posteriori* probability (MAP) rule, i.e.

$$\alpha = \log \frac{P(\omega_j)}{P(\omega_i)} \tag{17}$$

and d_{ij} is the Mahalanobis distance between the two classes ω_i and ω_j, and is given by (Tou and Gonzalez 1974, Richards 1993):

$$d_{ij} = (\boldsymbol{m_i} - \boldsymbol{m_j})^t \Sigma_{ij}^{-1} (\boldsymbol{m_i} - \boldsymbol{m_j}) \tag{18}$$

As d_{ij} in equation (16) depends on the feature considered, once the number of features to be selected has been fixed, one can perform feature selection by minimizing (16), used as a criterion function. In order to simplify the computation, one can

consider the upper bound to equation (16) provided by:

$$e_{ij} = [P(\omega_i) + P(\omega_j)]Q\left(\frac{\sqrt{d_{ij}}}{2}\right) \geqslant P_e(\omega_i, \omega_j) \tag{19}$$

which corresponds to fixing the threshold at the middle point of the Mahalanobis distance between the two classes, instead of using the optimal threshold computed according to the MAP rule.

Equations (16) and (19) can be used only in a two-class case. In order to choose from the literature (Lainiotis 1969, Hellman and Raviv 1970, Lainiotis and Park 1971, Devijver 1974, Garber and Djouadi 1988) a suitable upper bound to the error probability in multiclass cases for the purpose of our work, we evaluated two opposite requirements: the tightness of the bound to the error probability and the load for the computation of this bound. The former is a desirable property for an upper bound; however, bounds tightly related to the error probability are usually too complex to use (Devijver 1974). Therefore, in order not to increase the computational complexity too much, we selected a simple upper bound. It is provided by the sum of pairwise errors, computed for all pairs of classes as (Devijver 1974):

$$E_1 = \sum_{i=1}^{c} \sum_{j>1}^{c} P_e(\omega_i, \omega_j) \geqslant P_e \tag{20}$$

where $P_e(\omega_i, \omega_j)$ can be computed by equation (16). If we consider the pairwise upper bounds (19) instead of the errors (16), we can also write:

$$E_2 = \sum_{i=1}^{c} \sum_{j>1}^{c} e_{ij} \geqslant P_e \tag{21}$$

Feature selection can be performed according to the minimization of E_1 or E_2. The use of E_1 guarantees a better approximation for the error probability; on the other hand, the use of E_2 slightly reduces the computational load.

Equations (16) and (19) have been derived under the hypothesis of classes with equal covariance matrices. However, in practical cases, covariance matrices may be different. Therefore, for each pair of classes, we empirically compute Σ_{ij}, which appears in equation (18), as the mean value of the two covariance matrices Σ_i and Σ_j (Fukunaga 1990), i.e.:

$$\Sigma_{ij} = \frac{(\Sigma_i + \Sigma_j)}{2} \tag{22}$$

4. Procedure and data used for performance evaluation

4.1. *Performance evaluation procedure*

In order to compare the different criteria discussed in §2 with one another and with the proposed one, we utilized the classification accuracy obtained by giving the features (selected by the different algorithms) as input to the Bayes classifier (i.e. the classifier based on the Bayes rule for minimum error). To better isolate the effects of the various criterion functions on classification accuracy, we adopted an optimal search algorithm (otherwise, the result of the comparison might depend on the adopted suboptimal search algorithm). Consequently, we had to choose datasets with a moderate number of features so as to make optimal search algorithms applicable. In particular, we used the Branch and Bound algorithm (Fukunaga 1990,

Jain and Zongker 1997), which makes it possible to obtain optimal solutions in a reduced computation time, as compared with an exhaustive search.

We considered two different remote sensing datasets related to two agricultural areas, which are located in the 'Val Tiberina' (Italy) and near the village of 'Feltwell' (UK). For both datasets, the same kinds of experiments were carried out. In particular, each feature selection technique was used to select, from among n available features, the optimal subset of m features, for $m = 1, 2, \ldots, n-1$. It is worth noting that, as our experiments aimed at evaluating the effectiveness of feature selection criteria, they were carried out on training sets, without considering any test set, as is usually done for this kind of problem. In fact, the evaluation of the effectiveness of each feature set on the test set would also be influenced by the generalization capabilities of the features and not only by the effectiveness of each separability index. In our experiments, we assumed that data were Gaussian distributed. For the evaluation of the proposed criterion, we utilized E_2, which is computationally less expensive than E_1.

4.2. *Description of the dataset related to the 'Val Tiberina'*

The study area is part of the upper Tiber River Valley, delimited by the mountain chain of the Apennines to the east and by the subapennine hills to the west. A set of multitemporal remote sensing images, acquired by the Thematic Mapper (TM) sensor of the Landsat satellite (Richards 1993), were considered (figure 1). For our

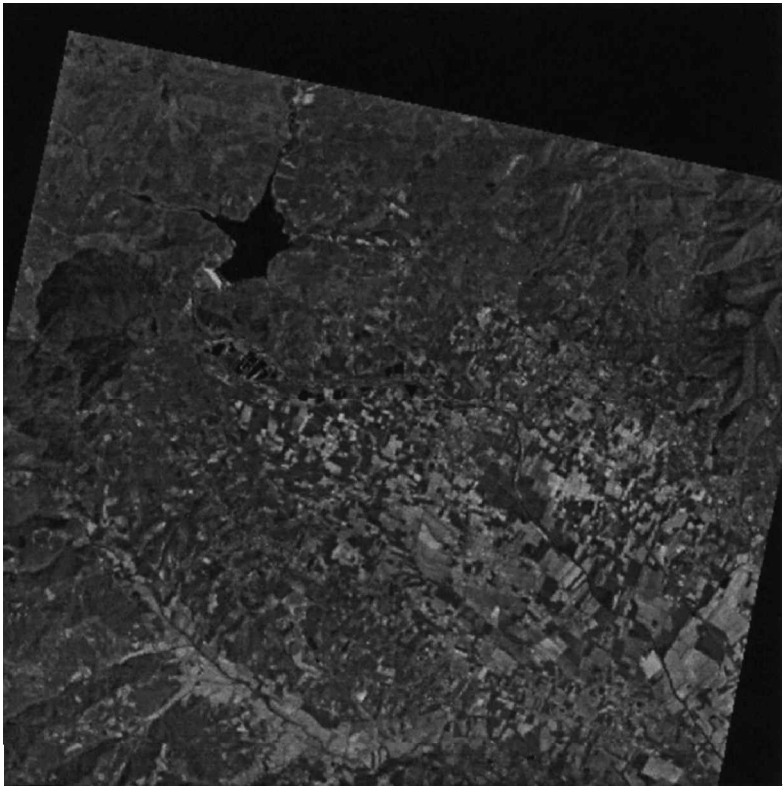

Figure 1. Image of the 'Val Tiberina' test site: channel 4 of the TM sensor.

experiments, two multispectral images (512×512 pixels) related to the same area were selected from two scenes taken on 16 March and 17 August 1991, respectively. Between the two acquisition dates, no change took place in the land cover of the area considered. The six non-thermal bands were chosen for each multispectral image. The analysis was carried out on a pixel basis, i.e. each pixel was considered as a pattern and characterized by a vector of 12 features. This feature vector was obtained by 'stacking' the feature vectors related to the images acquired at the single times. It is worth noting that the use of multitemporal features may improve the capability of the resulting set of features for distinguishing among different land-cover classes (some classes are separable by using features related to the first date, other classes are separable by using features related to the second date). From the available ground data, we selected the following four agricultural classes to be used in our experiments: wheat, corn, sunflowers and tobacco. A training set composed of 4592 samples was generated and utilized for the experiments (see table 1).

4.3. *Description of the 'Feltwell' dataset*

A section (250×350 pixels) of an image acquired by a multispectral scanner installed on an airplane (i.e. a Daedalus 1268 Airborne Thematic Mapper (ATM) scanner (Richards 1993)) was considered (figure 2). The flight took place in July 1989. In order to characterize each pixel, the six spectral bands corresponding to the TM channels (with the exception of the thermal channel) were employed. In addition, 11 nonlinear combinations of spectral channels (the so-called 'vegetation indexes' (Swain and Davis 1978, Richards 1993)) were used. According to the common denominations of the TM channels (Richards 1993), the following combinations were considered: (TM4− TM3)/(TM4+TM3), (TM5− TM4)/(TM5+TM4), TM3/TM1, TM3/TM2, TM4/TM1, TM4/TM2, TM4/TM3, TM4/TM5, TM4/TM7, TM5/TM1 and TM7/TM3. As a result, each pixel was characterized by a vector of 17 features. The use of the aforementioned vegetation indexes introduces redundant information into the feature set. Consequently, it makes the tests of the considered feature-selection criteria more interesting. A training set composed of 1431 samples belonging to five different agricultural classes was generated and utilized for the experiments (see table 2). The considered agricultural classes were wheat, sugar beets, potatoes, carrots and stubble.

5. Experimental results

5.1. *Results on the 'Val Tiberina' dataset*

Figure 3 shows the behaviour of the classification accuracies obtained by applying the Bayesian classifier to the features extracted by the considered criteria. In particular, the accuracies have been plotted versus the numbers of selected features. For the

Table 1. Classes and related numbers of pixels in the considered Val Tiberina training set.

Class	Number of pixels
Wheat	2129
Corn	1274
Sunflowers	589
Tobacco	600

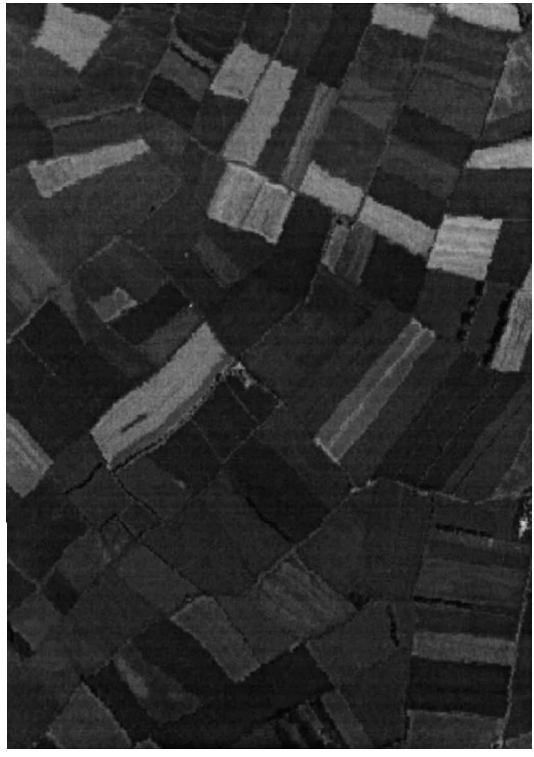

Figure 2. Image of the 'Feltwell' test site: channel 9 of the ATM sensor.

Table 2. Classes and related numbers of pixels in the considered Feltwell training set.

Class	Number of pixels
Wheat	283
Sugar beets	583
Potatoes	156
Carrots	256
Stubble	153

computation of the classification accuracy, the percentage of correctly classified pixels, with respect to the total number of pixels in the training set, has been utilized. From the behaviours given in the diagram, it is easy to deduce that, for this dataset, it is not useful to consider more than seven features. With seven features, the classification accuracy curve becomes flat for most of the criteria (i.e. addition of further features does not significantly improve classification accuracy). The analysis of figure 3 points out how the proposed criterion function (i.e. E_2) allows us to obtain an accuracy that is often higher than the ones reached by using the features selected by the other criteria. To better evaluate the obtained results, figure 4 presents the histograms of the average classification accuracies computed for all the considered techniques. The average accuracies were computed by summing the accuracies obtained by using the selected subsets of features (with the numbers of features ranging between 1 and 7) and by dividing the sum total by 7. The best results were

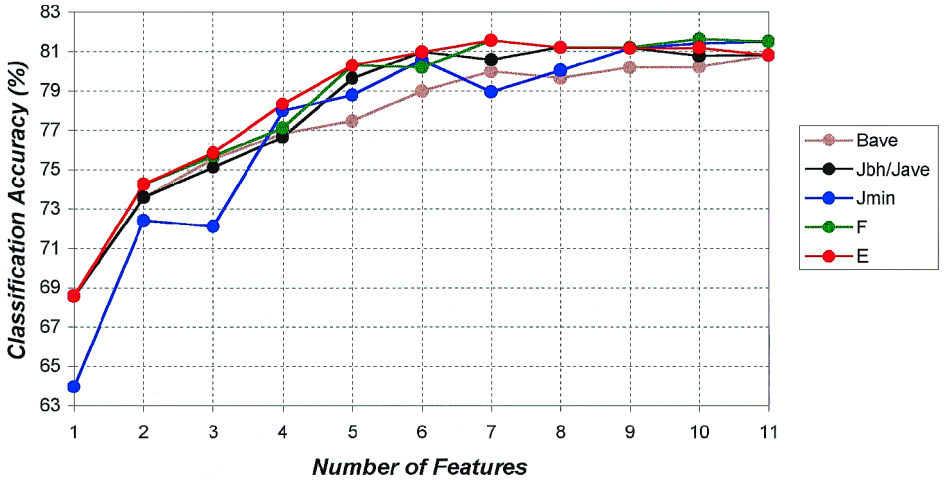

Figure 3. Comparison of the classification accuracies provided by the considered techniques for the 'Val Tiberina' dataset. (B_{ave} = average Bhattacharyya distance; J_{ave} = average J–M distance; J_{bh} = J–M distance generalized according to the Bhattacharyya bound to the Bayes error; J_{min} = J–M distance between the least distinguishable pair of classes; F = index based on scatter matrices; E = proposed criterion.)

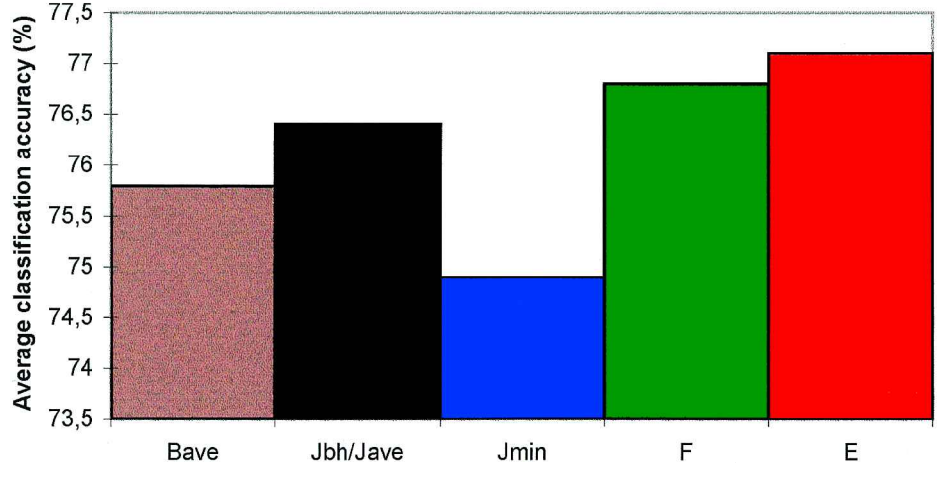

Figure 4. Average classification accuracies provided by the considered techniques for the 'Val Tiberina' dataset (the averages were computed by using numbers of selected features ranging between 1 and 7).

obtained by our criterion. It performed slightly better than the criterion based on scatter matrices, which turned out to be the best of the classical criteria. Average accuracies close to those obtained by the F index were obtained by J_{bh} and J_{ave}, which always selected the same subsets of features from the aforesaid dataset. The performances of B_{ave} and J_{min} were worse.

5.2. *Results on the 'Feltwell' dataset*

Figure 5 shows the behaviour of the classification accuracies obtained by the different algorithms versus the number of selected features. For this dataset, in order

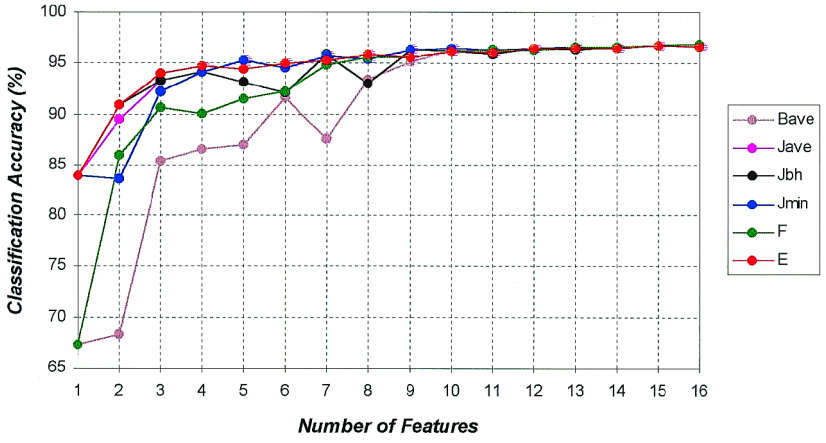

Figure 5. Comparison of the classification accuracies provided by the different techniques for the 'Feltwell' dataset.

to reach the saturating effect of the classification accuracy, eight can be regarded as a suitable number of features to be provided as input to the classifier. Therefore, in order to compare the performances of the different techniques, each average classification accuracy was computed considering a number of selected features from one to eight (figure 6). As was the case with the 'Val Tiberina' dataset, the proposed criterion provided the best average accuracy. It performed slightly better than the J_{bh} index, which gave the best average accuracy among the classical techniques. J_{ave} and J_{min} provided average accuracies close to that of J_{bh}. By contrast, F and B_{ave} yielded definitely worse performances.

5.3. *Discussion of results*

A global analysis of the obtained results suggests that, for both datasets, the proposed criterion turned out to be the most effective, as the subsets of features selected according to it allowed us to obtain the highest average classification

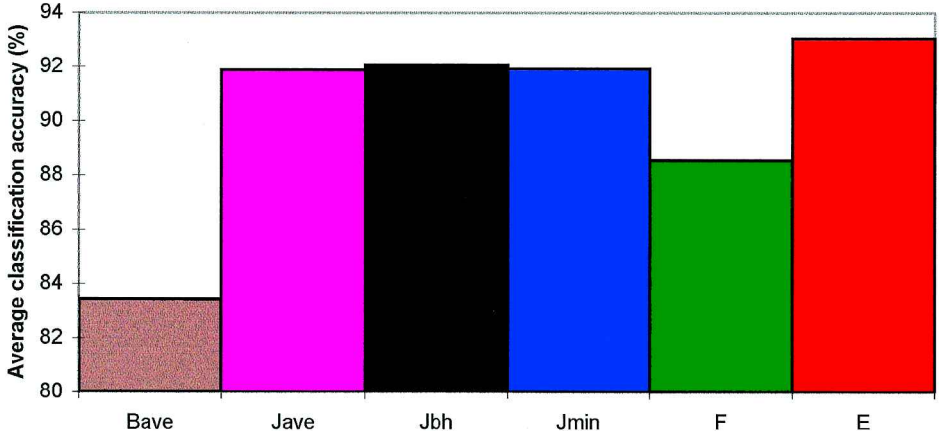

Figure 6. Average classification accuracies provided by the considered techniques for the 'Feltwell' dataset (the averages were computed by using numbers of selected features ranging between 1 and 8).

accuracy. Concerning the classical criteria, it is possible to state that J_{bh} and J_{ave} behaved rather well, providing average classification accuracies that were slightly worse than those attained by the proposed approach. The F index behaved in a different way on the two datasets: on the 'Val Tiberina' dataset, it provided an average classification accuracy that was very close to the best one, whereas, on the 'Feltwell' dataset, it reached an unsatisfactory average classification accuracy. The J_{min} index exhibited behaviours opposite to the ones of the F index on the two datasets. In particular, it provided an acceptable average classification accuracy only on the 'Feltwell' dataset. The criterion based on B_{ave} provided low average accuracies on both datasets.

In our opinion, even if, in some cases, the above results may appear contradictory, they can be interpreted as a confirmation that the critical point concerns the way in which the considered criteria combine the separability of each pair of classes in the computation of the global multiclass separability measure. Therefore, satisfactory results were obtained, on both datasets, by applying the three considered criteria (i.e. E_2, J_{ave} and J_{bh}) that take into account the separability of all couples of classes and weight them in a similar way, as they influence the overall classification error.

The best choice between the two generalizations of the J–M distance, J_{bh} and J_{ave}, in a general case, depends on the importance assigned to minority classes, as J_{bh} gives a greater weight to classes with low *a priori* probabilities, as compared with J_{ave} (Bruzzone *et al.* 1995). Concerning J_{min}, results confirm that considering only the couple of least separable classes may cause the selection of a feature subset that is globally not effective.

The low performances of the Bhattacharyya distance prove what is already well known in the literature, i.e. the non-saturating behaviour of this distance limits its effectiveness in feature-selection problems involving more than two classes.

Finally, the fact that the F index considers the distances of classes from the overall distribution of all samples (and not pairwise class distances, see equation (10)) and the lack of saturation effect for increasing distances may also explain the low performances of this criterion on the 'Feltwell' dataset. Therefore, it may be considered less reliable and not suited to all situations.

6. Conclusions

In this paper, we have focused attention on the feature-selection problem in multiclass cases. We have considered and compared various classical feature-selection criteria with one another and with the criterion proposed by us. In particular, for one of the considered classical criteria (i.e. the J–M distance), we have analysed three generalizations of two-class cases to multiclass ones. Experiments on two remote sensing datasets have been described.

For both datasets, the proposed criterion turned out to be particularly effective, as the subset of features it selected allowed us to obtain the highest average classification accuracy. Several experiments, not reported in this paper, were carried out to evaluate the accuracies provided by the features selected by the other formulation (i.e. E_1) of the proposed technique. However, no improvements were obtained over the classification accuracies reached by the feature subsets selected by E_2.

The proposed criterion has been formulated for data characterized by classes with Gaussian distributions, which can be considered a reasonable assumption in remote sensing applications involving the use of passive sensors. It is worth noting, however, that the proposed criterion performed well also when we used the dataset

including vegetation indexes, for which the assumption of Gaussian distributions of classes does not hold.

It is worth providing some information about the computational loads required by the different criteria. Neglecting the aspects related to the search algorithms (as they are beyond the scope of this work), it is interesting to note the different computational loads required to calculate the functionals for the different criteria considered. After fixing the number of features, the criteria can be compared by considering the matrix computations (which are the most time-consuming) needed by the different algorithms. It is easy to observe that the proposed criterion E_2 requires only the computation of one inverse matrix for each couple of classes (see equations (18) and (19)), whereas the B_{ave} criterion and all the considered J–M criteria require also the computations of one determinant for each class and of one determinant for each couple of classes (see equations (3) and (6)). Only the computation of the F index is simpler than that of the criterion function E_2 because it requires the calculation of one determinant of two global matrices instead of the calculation of several determinants for each class or couple of classes (see equation (12)). By the contrast, the F index seems to be less reliable, as was confirmed in one of our experiments.

The results of the described investigation can also be useful in selecting the bands of hyperspectral remote sensing images, as all the considered criterion functions can be applied to such images. However, given the very large number of bands involved, the resulting huge computational complexity makes it mandatory to adopt suboptimal search algorithms.

Although multiclass criteria based on error bounds have been presented in the pattern recognition literature, such criteria are not usually utilized by the remote sensing community. The criterion proposed in this paper is based on an error bound and, from a computational viewpoint, is less expensive than classical criteria such as those based on the Bhattacharyya distance and on the J–M distance. In addition, it proved effective for the real remote sensing data considered in our experiments. Therefore, in our opinion, it can be regarded as a valid alternative to other classical criteria commonly applied to remote sensing data acquired by passive sensors.

Acknowledgments

This research was conducted within the framework of the research project 'Sviluppo di metodi integrati di classificazione agroecologica tramite dati di telerilevamento per la gestione delle risorse naturali', supported by the Italian Space Agency.

The authors wish to thank CNR- IATA (Florence, Italy) and Hunting Technical Services Ltd. (UK) for providing the images of the 'Val Tiberina' and 'Feltwell' areas, respectively.

References

Bruzzone, L., Roli, F., and Serpico, S. B., 1995, An extension of the Jeffreys–Matusita distance to multiclass cases for feature selection. *IEEE Transactions on Geoscience and Remote Sensing*, **33**, 1318–1321.

Devijver, P. A., 1974, On a new class of bounds on Bayes risk in multihypothesis pattern recognition. *IEEE Transactions on Computer*, **23**, 70–80.

Fukunaga, K., 1990, *Introduction to Statistical Pattern Recognition*, 2nd edn (New York: Academic Press).

Garber, F. D., and Djouadi, A., 1988, Bounds on the bayes classification error based on

pairwise risk functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **10**, 281–288.

Hellman, M. E., and Raviv, J., 1970, Probability of error, equivocation, and the Chernoff bound. *IEEE Transactions on Information Theory*, **IT-16**, 368–372.

Jain, A., and Zongker, D., 1997, Feature selection: evaluation, application and small sample performances. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**, 153–158.

Kailath, T., 1967, The divergence and the Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, **15**, 52–60.

Krishnan, S., Samudravijaya, K., and Rao, P. V. S., 1996, Feature selection for pattern classification with gaussian mixture models: a new objective criterion. *Pattern Recognition Letters*, **17**, 803–809.

Lainiotis, D. G., 1969, A class of upper bounds on probability of error for multihypotheses pattern recognition. *IEEE Transactions on Information Theory*, **IT-15**, 730–731.

Lainiotis, D. G., and Park, S. K., 1971, Probability of error bounds. *IEEE Transactions on Systems, Man, and Cybernetics*, 175–178.

Lee, C., and Landgrebe, D. A., 1993, Feature extraction based on decision boundaries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15**, 388–400.

Liu, S. S., and Jernigan, M. E., 1990, Texture analysis and discrimination in additive noise. *Computer Vision, and Image Processing*, **49**, 52–67.

Mausel, P. W., Kramber, W. J., and Lee, J. K., 1990, Optimum band selection for supervised classification of multispectral data. *Photogrammetric Engineering and Remote Sensing*, **56**, 55–60.

Novovicova, J., Pudil, P., and Kittler, J., 1996, Divergence based feature selection for multimodal class densities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**, 218–223.

Pudil, P, Novovicova, J., and Kittler, J., 1994, Floating search methods in feature selection. *Pattern Recognition Letters*, **15**, 1119–1125.

Richards, J. A., 1993, *Remote Sensing: Digital Image Analysis*, 2nd edn (New York: Springer).

Schowengerdt, R. A., 1997, *Remote Sensing: Models and Methods for Image Processing*, 2nd edn (New York: Academic Press).

Siedlecki, W., and Sklansky, J., 1989, A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, **10**, 335–347.

Swain, P. H., and Davis, S. M., 1978, *Remote Sensing: The Quantitative Approach* (New York: McGraw-Hill).

Thomas, I. L., Ching, N. P., Benning, V. M., and D'Aguanno, J. A., 1987, A review of multi-channel indices of class separability. *International Journal of Remote Sensing*, **8**, 331–350.

Tou, J. T., and Gonzalez, R. C., 1974, *Pattern Recognition Principles* (London: Addison-Wesley).