An Approach to Feature Selection and Classification of Remote Sensing Images Based on the Bayes Rule for Minimum Cost

Lorenzo Bruzzone, *Member, IEEE*

***Abstract*—Classification of remote-sensing images is usually carried out by using approaches aimed at minimizing the overall error affecting land-cover maps. However, in several remote-sensing problems, it could be useful to perform classification by taking into account the different consequences (and hence the different costs) associated with each kind of error. This allows one to obtain land-cover maps in which the total classification cost involved by errors is minimized, instead of the overall classification error. To this end, in this paper, an approach to feature selection and classification of remote-sensing images based on the Bayes rule for minimum cost (BRMC) is proposed. In particular, a feature-selection criterion function is presented that permits one to select the features to be given as input to a classifier by taking into account the different cost associated with each confused pair of land-cover classes. Moreover, a classification technique based on the BRMC and implemented by using a neural network is described. The results of experiments carried out on a multisource data set concerning the Island of Elba (Italy) point out the ability of the proposed minimum cost approach to produce land-cover maps in which the consequences of each kind of error are considered.**

***Index Terms*—Bayes rule for minimum cost, feature selection, image classification, remote sensing, risk assessment.**

I. Introduction

PRODUCTION of land-cover maps by using automatic classification techniques is one of the main applications of remotely sensed images [1]. Such maps can be utilized for various purposes. In recent years, ever-increasing attention has been devoted to employing land-cover maps to define policies for environmental interventions based on the spatial distribution of different land covers in the area considered. In some cases, these interventions aim to reduce the risks of natural disasters (e.g., forest fires, floods, etc.) in areas with land covers characterized by high risk levels. In other cases, the purpose of interventions is to realize infrastructures on the basis of the land-covers in the regions selected. In all such cases, different kinds of errors (i.e., different confused pairs of classes) may lead to different wrong interventions on the land that may result in consequences of different severity. Therefore, each kind of error can be associated with a more or less high cost, depending on the implications of the error. Such a cost should be taken into account in the classification process in order to minimize the errors that cause the most severe consequences to a given application.

In the literature, several methodologies for automatic classification of remote-sensing images used to produce land-cover maps have been presented [1]–[9]. These methodologies (e.g., neural networks [2]–[5], fuzzy logic [7], [8], and knowledge-based paradigms [8], [9]) face various aspects of automatic classification. Many of the proposed approaches are based directly or indirectly on the Bayes rule for minimum error (BRME), which aims at a classification that is affected by a minimum overall error [10]. On the contrary, little attention has been devoted to the development of approaches to performing a classification that takes into account the different cost that each kind of error involves for the application considered [11], [12]. However, in some cases, it may be more appropriate to design a classification system aimed at minimizing errors that result in high costs, rather than minimizing the overall classification error. To this end, an approach based on the Bayes rule for minimum cost (BRMC) [10], which aims to minimize the total classification cost, should be adopted.

In this paper, an approach to feature selection and classification based on the BRMC is presented as a valid alternative to the approaches based on the BRME for remote-sensing applications, in which different kinds of errors result in different costs. In particular, a feature-selection criterion function is proposed that selects effective subsets of features to be given as input to a classifier by taking into account the cost associated with each confused pair of land-cover classes. In addition, a classification technique that implements the BRMC by using a neural network is presented. Such a classification technique, which is of the nonparametric type, is suitable to process multisource data. Experiments are described concerning the problem of obtaining land-cover maps suited to generating risk maps of forest fires by using both remote-sensing images and ancillary data. The results of these experiments confirm the validity of the proposed approach.

The paper is organized into six sections. The BRME and the BRMC are briefly defined in Section II. In Section III, a feature-selection criterion function that takes into account the cost of each error is proposed. A neural-network classification technique based on the BRMC is presented in Section IV. In Section V, the data set used for experiments, the preprocessing applied to data, and the strategy adopted to define the cost matrix are described. Moreover, experimental results are reported and discussed. Finally, conclusions are drawn in Section VI.

Manuscript received May 18, 1998; revised April 23, 1999. This work was supported in part by the Italian Space Agency, Ispra, and the European Commission (Environment Programme II, Contract EV5V-CT94-0481).

The author is with the Department of Biophysical and Electronic Engineering, University of Genoa, Genova, Italy (e-mail: lore@dibe.unige.it).

Publisher Item Identifier S 0196-2892(00)00411-3.

II. Bayes Rule for Minimum Error (BRME) and Bayes Rule for Minimum Cost (BRMC)

Let us consider a remote-sensing image in which a generic pixel, described by an n-dimensional feature vector $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ in the feature space X, is to be assigned to one of p different land-cover classes $\Omega = (\omega_1, \omega_2, \cdots, \omega_p)$ characterized by the *a priori* probabilities $P(\omega_i)$, $\omega_i \in \Omega$. Let $p(\boldsymbol{x}/\omega_i)$ be the conditional density function for the feature vector $\boldsymbol{x}$ given the class $\omega_i \in \Omega$. Let C be the cost matrix of dimension $p \times p$ in which each element c_{ij} represents the cost of deciding on $\boldsymbol{x} \in \omega_i$ when $\boldsymbol{x} \in \omega_j$ (the decisions of the classifier are given in the matrix rows, while the true classes are given in the columns). A high cost c_{ij} corresponds to a situation in which the confusion of the class ω_j with the class ω_i is very critical.

It is well known that a classifier based on the BRME assigns the pixel characterized by the feature vector $\boldsymbol{x}$ to the class ω_k if the posterior probability $P(\omega_k/\boldsymbol{x})$ is the highest one [10]

$$\boldsymbol{x} \in \omega_k, \qquad \text{if } P(\omega_k/\boldsymbol{x}) = \max_{\omega_i \in \Omega} \{P(\omega_i/\boldsymbol{x})\}. \tag{1}$$

However, a classifier based on the BRMC associates the pixel described by the feature vector $\boldsymbol{x}$ with the class ω_k if the conditional cost $R(\omega_k/\boldsymbol{x})$ is the lowest one [10],

$$\boldsymbol{x} \in \omega_k \qquad \text{if } R(\omega_k/\boldsymbol{x}) = \min_{\omega_i \in \Omega} \{R(\omega_i/\boldsymbol{x})\} \tag{2}$$

where $R(\omega_i/\boldsymbol{x})$ is defined as [10]

$$\begin{aligned} R(\omega_i/\boldsymbol{x}) &= c_{i1}P(\omega_1/\boldsymbol{x}) + c_{i2}P(\omega_2/\boldsymbol{x}) + \cdots \\ &\quad + c_{ip}P(\omega_p/\boldsymbol{x}), \qquad \forall\, \omega_i \in \Omega. \end{aligned} \tag{3}$$

The classification carried out by using the BRME minimizes the overall error expressed as [10]

$$\begin{aligned} E = \int_x \min[&P(\omega_1)p(\boldsymbol{x}/\omega_1),\, P(\omega_2)p(\boldsymbol{x}/\omega_2),\, \cdots, \\ &P(\omega_p)p(\boldsymbol{x}/\omega_p)]\, dx \end{aligned} \tag{4}$$

while the classification carried out by using the BRMC minimizes the total cost expressed as [10]

$$R = \int_x \min[R(\omega_1/\boldsymbol{x}),\, R(\omega_2/\boldsymbol{x}),\, \cdots,\, R(\omega_p/\boldsymbol{x})]\, dx. \tag{5}$$

Consequently, the methods based on the BRMC usually yield a classification affected by a lower total cost and a higher overall error, as compared with the classification provided by the methods based on the BRME.

III. Feature-Selection Criterion Function Based on the BRMC

The choice of a set of features that can best discriminate among land-cover classes to be recognized by a classifier is one of the main problems involved in the development of a classification system. In remote sensing, besides the features related to the spectral channels acquired by sensors, other features extracted by the processing of the information contained in these spectral channels (e.g., texture features [13]) or related to ancillary data [14] are often considered. Even though these features may increase the capability to distinguish land-cover classes, the resulting feature set often contains redundant information. Consequently, in the phase of the system design, it is recommended that only the most effective features from the set of available ones be selected and that the redundant ones be discarded. From a more formal point of view, feature selection aims at choosing a subset of m features from among n available ones (with $m < n$) that provide the best separation of land-cover classes in the feature space. The reduction in the number of features results in a decrease in the computational time taken by the processing system, thanks to the lower computational loads of the feature-extraction and classification tasks. Moreover, in practical situations involving a limited number of training samples, a reduction in the number of features may also increase classification accuracy (Hughes phenomenon) [10].

Feature-selection techniques usually involve a criterion function and a search algorithm. The former aims at evaluating the separability of classes for a given subset of features. The latter identifies the subset of features that maximize the adopted criterion function. In this paper, the focus is on criterion functions, with particular emphasis on the definition of a criterion function that evaluates the degree of effectiveness of features by taking into account the costs associated with errors. Search algorithms are not considered here, because any classical search algorithm (e.g., branch and bound [10], [15], sequential forward-floating selection [16], etc.) can be applied to the proposed criterion function.

Criterion functions are usually based on separability indexes that express the effectiveness of each subset of features. Several separability indexes have been proposed in the remote-sensing literature [1], [17]–[20]. When more than two classes are considered, these indexes are generally based on an average distance among classes d_{ave} defined as [1], [17]

$$d_{ave} = \sum_{i=1}^{p} \sum_{j=1}^{p} P(\omega_i)P(\omega_j)\, d_{ij} \tag{6}$$

where d_{ij} is a statistical distance between the pair of classes ω_i and ω_j and depends on the set of features considered. Many statistical distances d_{ij} have been used in remote-sensing problems (e.g., Euclidean distance, divergence, transformed divergence, Bhattacharyya distance, and Jeffreys-Matusita distance [17]–[21]).

Criterion functions based on d_{ave}-weight pairwise distances without taking into account the costs associated with classes. Therefore, they are effective to select features that are suitable to minimize the overall classification error, but they are not appropriate to select features suitable to minimize the total classification cost. As a consequence, the use of these criterion functions may lead to the choice of features that are not effective in distinguishing classes that, if confused, involve high costs. For this reason, the criterion functions based on the average distance should be reformulated on the basis of the BRMC in order to make them able to select features by taking into account the costs associated with each confused pair of classes.

In order to derive a minimum-cost formulation of d_{ave}, it is useful to analyze the differences between the BRME and the BRMC. In particular, to understand how one can weight the pairwise distances d_{ij} by considering the cost related to each confused pair of classes, let us consider the aforementioned decision rules in the case of only two classes ω_1 and ω_2. In this case, the BRME [see (1)] can be rewritten as [10]

$$P(\omega_1)p(\boldsymbol{x}/\omega_1) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} P(\omega_2)p(\boldsymbol{x}/\omega_2). \tag{7}$$

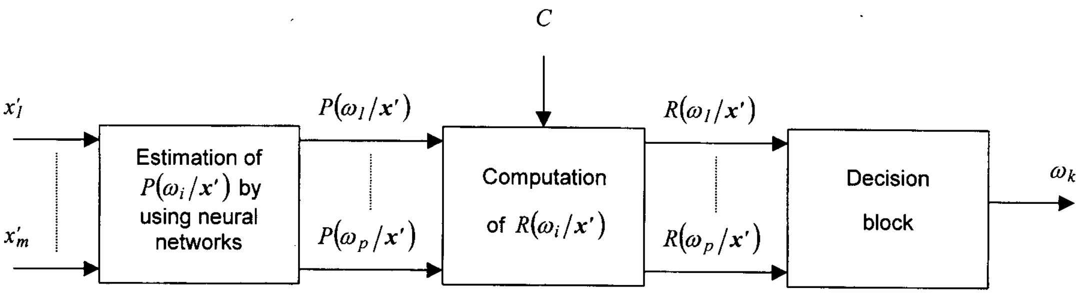

Fig. 1. Block diagram of the proposed classification technique.

Similarly, it is possible to prove that the BRMC [see (2)] can be rewritten as [10]

$$(c_{21} - c_{11})P(\omega_1)p(\boldsymbol{x}/\omega_1) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} (c_{12} - c_{22})P(\omega_2)p(\boldsymbol{x}/\omega_2). \quad (8)$$

By comparing (8) with (7), one can observe that the use of the BRMC is equivalent to "modifying" the *a priori* probabilities of classes in the BRME. In fact, the BRMC weights the conditional density functions of classes by using costs in addition to *a priori* probabilities. This makes it possible to take into account that a wrong decision on one class may be more critical than that on another class. On the basis of this observation, in order to select features by taking into account the cost of each error, the following minimum-cost formulation of (6) is proposed:

$$d_{cost} = \sum_{i=1}^{p} \sum_{j=1}^{p} (c_{ij} - c_{jj})(c_{ji} - c_{ii})P(\omega_i)P(\omega_j)d_{ij}. \quad (9)$$

Given a fixed number m of features to be selected, the subset of features that maximizes d_{cost}, as compared to the one that maximizes d_{ave}, provides a better separation of classes that would result in high costs if misclassified. This is a consequence of the fact that the importance given to the separability of each pair of classes obtained by using d_{cost} depends not only on the *a priori* probabilities of both classes but also on the cost entailed by the confusion of the two classes.

The application of a feature selection based on d_{cost} requires the definition of the cost matrix C, the estimation of the *a priori* probabilities $P(\omega_i)$ of the classes, and the computation of the distance d_{ij} between each pair of classes. The definition of the cost matrix C is a very critical step. It should be performed in close cooperation with experts in the considered application (or with end-users) who well realize the consequences of each type of error in practical situations. Concerning *a priori* probabilities, they can be estimated, as usually done in remote-sensing applications, on the basis of the frequency of each class in the training set [22]. Finally, distance computation depends on the kind of separability index adopted (e.g., divergence [17], Jeffreys–Matusita distance [17], [20]).

IV. Neural-Network Classification Technique Based on the BRMC

Let us assume that, at the end of the feature-selection phase, a generic pixel is described by an m-dimensional feature vector $\mathbf{x}' = (x'_1, x'_2, \cdots, x'_m)$ composed of the selected features only. Such features have to be given as input to a classifier to produce a land-cover map.

Classification based on the BRMC is performed by assigning each pixel to the land-cover class ω_k, for which the estimated conditional cost is the lowest [see (2)]. Given the cost matrix C, it is therefore necessary for each pixel to be analyzed to estimate the posterior probabilities $P(\omega_i/\boldsymbol{x}')$, $\forall\, \omega_i \in \Omega$, in order to compute the conditional costs $R(\omega_i/\boldsymbol{x}')$, $\forall\, \omega_i \in \Omega$ [see (3)].

A block diagram of the proposed classification scheme is shown in Fig. 1. Such a diagram is composed of three blocks devoted to the estimation of the $P(\omega_i/\boldsymbol{x}')$, to the computation of the $R(\omega_i/\boldsymbol{x}')$, and to the final decision to be made on the classification.

1) Estimation of $P(\omega_i/\boldsymbol{x}')$: Several methods have been proposed in the literature to estimate the posterior probabilities $P(\omega_i/\boldsymbol{x}')$, $\forall \omega_i \in \Omega$ [10], [23]. In the proposed classification technique, such estimations are performed by using a neural network [24]–[26]. This choice has been made in order to develop a nonparametric classifier that can be used to process multisensor and multisource data [2], [4]. In particular, a multilayer-perceptron (MLP) neural network [26], [27] with a fully connected architecture composed of one input layer (with m neurons as input features), one hidden layer, and one output layer (with p neurons as classes) is considered (see Fig. 2). Each unit (or neuron) of the network is characterized by a sigmoidal activation function. In order to make the ith network output provide an approximation of the posterior probability $P(\omega_i/\boldsymbol{x}')$, the error backpropagation (EBP) learning algorithm and the mean square error (MSE), considered as a cost function, are used [24]–[27]. In particular, the MSE to be minimized is defined as [24]–[27]

$$E = \frac{1}{n \cdot p} \sum_{i=1}^{p} \sum_{l=1}^{n_i} \sum_{k=1}^{p} \left[t_k^i - o_k(x_l^i)\right]^2 \quad (10)$$

where n and n_i are the total number of samples and the number of samples of the class ω_i, respectively, in the training set. t_k^i is the target for the kth output of the network for the samples of the class ω_i. $o_k(x_l^i)$ is the kth output of the network when the lth sample of the class ω_i has been presented as input. To obtain approximations for the posterior probabilities of the classes, the sum of the outputs of the network trained by using the EBP algorithm and the MSE cost function should be normalized to 1 [24], [25].

2) Computation of $R(\omega_i/\boldsymbol{x}')$: On the basis of the considered cost matrix C and of the estimates of the posterior probabilities $P(\omega_i/\boldsymbol{x}')$, $\forall\, \omega_i \in \Omega$ (provided by the MLP neural network),

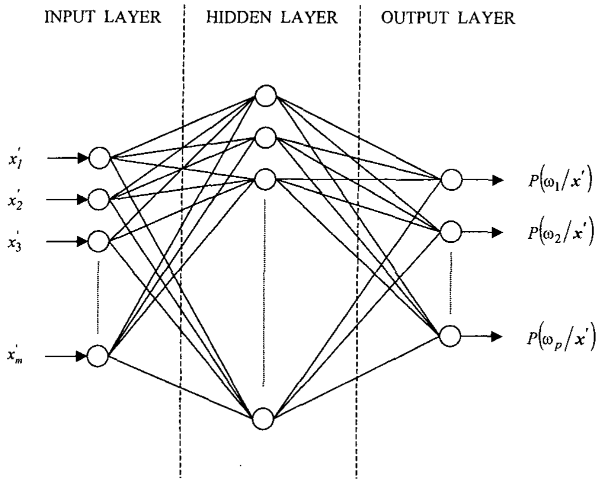

Fig. 2. Example of multilayer perceptron neural-network architecture for estimating posterior-class probabilities.

this block computes the conditional costs $R(\omega_i/\boldsymbol{x}')$, $\forall\,\omega_i \in \Omega$ according to (3).

3) Decision Block: The objective of this block is to make the final decision on the basis of the BRMC. In particular, each pixel to be classified is assigned to the class ω_k, for which the conditional cost is the minimum one. It is worth noting that this block plays a role similar to the one of the block that implements the "Winner-Takes-All" decision rule [28] in neural-network classifiers based on the BRME. However, in the present context, it selects the class corresponding to the lowest conditional cost instead of the class associated with the highest posterior probability.

V. Experimental Results

In order to give an example of application to a real case of the proposed feature-selection criterion function and classification technique based on the BRMC, the problem of achieving a land-cover map suited to being used to derive a risk map related to forest fires was addressed. The fire-risk map was obtained by assigning a risk value to each vegetation class present in the land-cover map, in accordance with suggestions of experts [29], [30].

The use of a classification approach based on the BRMC is particularly suitable for this application. In fact, different types of errors on the land-cover map (and hence on the risk map) may lead to different wrong interventions in the environment that may result in consequences of different gravity. In particular, the impact of a misclassification on the risk map depends on the risk values associated with the land covers confused by the classifier. On the one hand, it is not critical to confuse classes associated with similar risk levels. On the other hand, it is very critical to confuse classes associated with strongly different risk levels.

A. Data-Set Description

The considered data set refers to the western part of the Island of Elba, located in the Northern Tyrrhenian Sea (Italy). A section (400 × 326 pixels) of a scene acquired by the Thematic Mapper (TM) of the Landsat 5 satellite [1] was selected. The acquisition took place on August 25, 1992. Fig. 3 shows a false-color composition of channels 7, 5, and 4 of the TM sensor. The ground truth was used to prepare a thematic map of the considered section. Such a map was used as a reference map to assess both the classification errors and the related costs of errors. The analysis was carried out on a pixel basis (i.e., each pixel was considered as a pattern). For the experiments reported in this paper, the 16 land-cover classes listed in Table I were chosen. In all, 12 938 pixels were selected. Land-covers were randomly subdivided into two sets: 6472 training pixels were taken from one set and 6466 test pixels from the other (see Table I). The training set was used for feature selection and to train classifiers, and the test set was used for performance evaluations and comparisons.

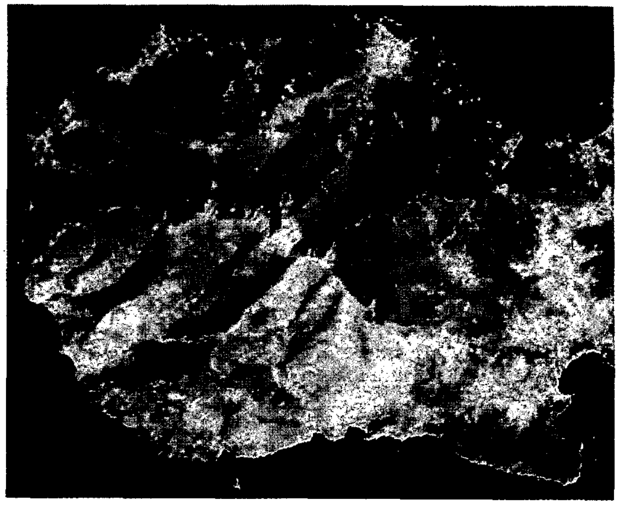

Fig. 3. Image of the test site: false-color composition of channels 7 (red), 5 (green), and 4 (blue) of the TM sensor.

B. Data Preprocessing

The considered TM images contain significant textural information that can be used to increase the separability of land-cover classes. In the literature, many techniques have been proposed to characterize remote-sensing image textures. For the present study, the texture features computed from the grey-level co-occurrence (GLC) matrix [31] were utilized. The GLC matrix constitutes a statistical approach to texture computation that has been successfully tested on remote-sensing images for land-cover mapping [13], [14]. In theory, the 12 GLC texture features proposed in [31] could be computed for each of the TM channels of the selected scene, thus obtaining a set of 84 texture features. In practice, in order to reduce the computational cost, the 12 texture features were computed for only one of the seven TM channels. In particular, a visual analysis of the images related to the different spectral channels highlighted that TM channel 5 (i.e., an infrared channel) contained the largest amount of textural information. Therefore, the GLC texture features were computed by using such a channel. The computation of the GLC texture features required the choice of a given number of parameter values for the computation of the GLC matrix (i.e., interpixel distance, window size, and orientation). Taking into account the

TABLE I
CLASSES AND RELATED NUMBERS OF PIXELS IN THE CONSIDERED TRAINING AND TEST SETS. THE FIRE-RISK VALUE ASSOCIATED WITH EACH CLASS IS GIVEN

Label	Land-cover class	Number of pixels in the training set	Number of pixels in the test set	Risk value
ω_1	Chestnut forest	210	210	8
ω_2	Dense pine forest	213	214	5
ω_3	Thin pine forest	256	256	4
ω_4	Dense low maquis	508	508	3
ω_5	Thin low maquis	195	195	2
ω_6	Thin low maquis with rocks	263	263	2
ω_7	Dense coppice	89	88	6
ω_8	Mixed coppice-pine forest	1444	1443	7
ω_9	Thin coppice	245	245	5
ω_{10}	Dense high maquis	229	228	5
ω_{11}	Thin high maquis	176	176	4
ω_{12}	Reforested land	89	88	6
ω_{13}	Pasture-land with rocks	803	802	1
ω_{14}	Pasture-land	1291	1290	1
ω_{15}	Pasture-land with trees	327	326	1
ω_{16}	Urban and agricultural land	134	134	1
	Total	6472	6466	

fine textures of the considered TM images, the GLC matrix was computed by using an interpixel distance equal to 1 pixel and a window size of 9 × 9 pixels. The texture was assumed to be isotropic (a visual analysis did not reveal any particular dominant orientation), then it was computed for an angle of 0° only. The original 256 grey levels were mapped into 64 levels in order to reduce the time required by the computation of the GLC matrix and to make the estimates of the terms of the GLC matrix more reliable.

A digital terrain model (DTM) of the selected area was also considered [29]. In particular, the DTM was composed of three images containing the elevation, the slope, and the aspect, respectively, with a pixel size of 30 × 30 m. These images were georeferenced on the TM images.

In all, 21 features were chosen to form a feature vector for each pixel. In particular, the 6 TM channels in the visible and in the infrared spectrum (the thermal band was disregarded), and both the 12 GLC texture features and the three DTM features described in this section were considered. A smoothing filter (i.e., the mean filter [1]) was applied to all the TM images in order to reduce pixel-to-pixel intensity variations (a window size of 3 × 3 pixels was used for the filtering process). All the 21 features were normalized to a range between 0 and 1.

C. Definitions of Risk Values and of the Cost Matrix

In order to achieve a risk map related to the selected area, each land-cover class was associated with a specific value of forest-fire risk. In particular, eight different risk values were defined according to the suggestions of experts [29] (from 1, i.e. low risk, to 8, i.e. high risk) (see Table I). The cost matrix C used for the proposed feature-selection criterion function and classification technique was defined, taking into account the risk values associated with each pair of land-cover classes. In particular, the following strategy was adopted. If the risk value of one class ω_i is close to that of another class ω_j, then the resulting costs c_{ij} and c_{ji} are low. On the contrary, if two land-cover classes ω_i and ω_j have very different risk values, the costs c_{ij} and c_{ji} are high. In particular, the cost is higher if the class associated with the higher risk value is confused with the one associated with the lower risk value, whereas it is lower in the opposite case. In fact, in the first situation, errors do not lead to devising prevention strategies for areas with a high fire probability, whereas, in the second situation, errors do lead to the definition of prevention strategies for areas with a low fire probability. It is clear that the first kind of errors may result in more severe consequences. On the basis of the above strategy, the following procedure was adopted to define costs:

$$c_{ij} = \begin{cases} 0, & \text{if } \omega_i = \omega_j \\ \Delta\text{risk}_{ij} + 1, & \text{if } \omega_i \neq \omega_j \text{ and} \\ & \quad \text{risk}(\omega_i) \geq \text{risk}(\omega_j) \\ k(\Delta\text{risk}_{ij} + 1)^2, & \text{if } \omega_i \neq \omega_j \text{ and} \\ & \quad \text{risk}(\omega_i) < \text{risk}(\omega_j) \end{cases} \tag{11}$$

where $\text{risk}(\omega_i)$ is the risk value associated with the class ω_i (see Table I), and Δrisk_{ij} is defined as

$$\Delta\text{risk}_{ij} = |\text{risk}(\omega_j) - \text{risk}(\omega_i)|. \tag{12}$$

Here, k is a constant parameter that tunes the degree of difference between the cost of confusing one class associated with a high risk value, with another class associated with a low risk value and the cost of the opposite situation. In the experiments reported in this paper, $k = 1$ was chosen. The cost matrix obtained by using (11) is shown in Table II.

It is worth noting that the strategy adopted to choose costs represents only an example used to analyze the performances

TABLE II
COST MATRIX USED IN THE EXPERIMENTS. THE TRUE LAND-COVER CLASSES ARE GIVEN IN THE COLUMNS, AND THE DECISIONS OF THE CLASSIFIER ARE GIVEN IN THE ROWS

Decision	**True class**															
	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	ω_{11}	ω_{12}	ω_{13}	ω_{14}	ω_{15}	ω_{16}
ω_1	0	4	5	6	7	7	3	2	4	4	5	3	8	8	8	8
ω_2	16	0	2	3	4	4	4	9	1	1	2	4	5	5	5	5
ω_3	25	4	0	2	3	3	9	16	4	4	1	9	4	4	4	4
ω_4	36	9	4	0	2	2	16	25	9	9	4	16	3	3	3	3
ω_5	49	16	9	4	0	1	25	36	16	16	9	25	2	2	2	2
ω_6	49	16	9	4	1	0	25	36	16	16	9	25	2	2	2	2
ω_7	9	2	3	4	5	5	0	4	2	2	3	1	6	6	6	6
ω_8	4	3	4	5	6	6	2	0	3	3	4	2	7	7	7	7
ω_9	16	1	2	3	4	4	4	9	0	1	2	4	5	5	5	5
ω_{10}	16	1	2	3	4	4	4	9	1	0	2	4	5	5	5	5
ω_{11}	25	4	1	2	3	3	9	16	4	4	0	9	4	4	4	4
ω_{12}	9	2	3	4	5	5	1	4	2	2	3	0	6	6	6	6
ω_{13}	64	25	16	9	4	4	36	49	25	25	16	36	0	1	1	1
ω_{14}	64	25	16	9	4	4	36	49	25	25	16	36	1	0	1	1
ω_{15}	64	25	16	9	4	4	36	49	25	25	16	36	1	1	0	1
ω_{16}	64	25	16	9	4	4	36	49	25	25	16	36	1	1	1	0

of the proposed approach. Generally, the choice of costs and the relations among them should be carefully evaluated on the basis of a specific application and of end-user requirements.

D. Results and Discussion

Experiments were carried out to compare the effectiveness of the proposed feature-selection criterion function and classification technique based on the BRMC with that of the classical methods based on the BRME. To this end, the total classification cost and the overall classification error obtained by giving the features selected by using d_{cost} as input to the proposed neural classifier based on the BRMC were compared with the total cost and the overall error obtained by giving the features selected by using d_{ave} as input to a neural classifier based on the BRME. The total cost was computed as the sum of each cost c_{ij} multiplied by the number of related classification errors.

In the experiments carried out, the Jeffreys–Matusita distance [17], [20] was considered as the distance d_{ij} to perform feature selection by using the criterion functions given in (6) and (9). Such a distance is defined as [17]

$$d_{ij} = \left\{ \int_x \left[\sqrt{p(\boldsymbol{x}/\omega_i)} - \sqrt{p(\boldsymbol{x}/\omega_j)} \right]^2 d\boldsymbol{x} \right\}^{1/2}. \tag{13}$$

The general definition of the Jeffreys–Matusita distance does not require any particular assumption on the distributions of the conditional density functions of classes. For simplicity, this distance is computed here under the hypothesis of Gaussian distributions. It is worth noting that this is an approximation, because textural and DTM features may have distributions that do not accurately fit the Gaussian model. However, as such an approximation is used for both d_{cost} and d_{ave}, it does not affect the comparative evaluation of the two feature-selection criterion functions. The Branch and Bound search algorithm [10], [16] was chosen to identify the subsets of features that maximize the considered criterion functions.

Preliminary feature-selection trials were performed to find the number m of features to be given as input to the classifiers in all the carried out experiments. In particular, d_{cost} and d_{ave} were used to select the best subsets of k features (with $k = 1, \cdots n - 1$) from among the $n = 21$ available ones. Figs. 4 and 5 show the behaviors of d_{cost} and d_{ave}, respectively, versus the numbers of selected features. By analyzing the behaviors in these diagrams, one can see that, for more than nine features, the values of both criterion functions become flat (i.e., the addition of further features does not significantly increase the values of such functions). Consequently, nine features were selected to carry out the experiments described in this paper.

The nine best features selected by using d_{cost} are channels 1, 3, and 4 of the TM, five texture features (correlation, difference variance, information measure of correlation and sum average, and variance), and the elevation feature of the DTM. These nine features were given as input to a fully connected MLP with nine input neurons, 24 neurons in the hidden layer, and 16 output neurons, in order to estimate the $P(\omega_i/\boldsymbol{x}')$. The EBP learning procedure was used to train the network, which was initialized with random weights. As a convergence criterion, an MSE smaller than 0.015 was required. Classification was performed on the test set. A total cost equal to 6335 was obtained, with an overall classification error equal to 18.8%.

To assess the validity of the proposed classification approach, the total classification cost and the overall classification error related to the aforementioned experiment were compared with the total cost and the overall error resulting from giving the nine best

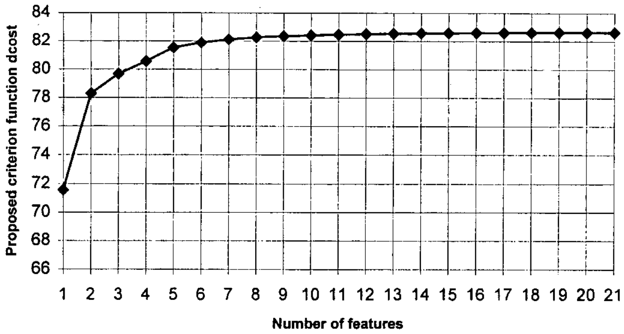

Fig. 4. Behavior of the proposed feature-selection criterion function d_{cost} versus the number of selected features.

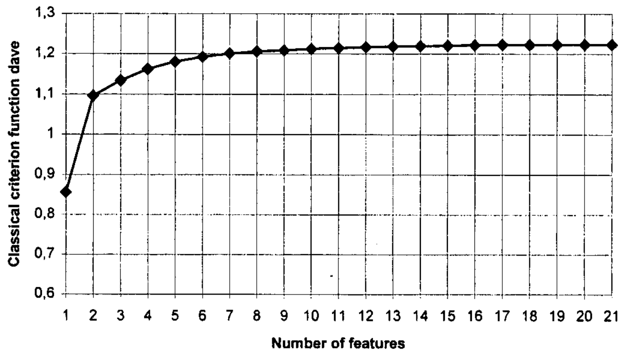

Fig. 5. Behavior of the classical feature-selection criterion function d_{ave} versus the number of selected features.

features selected by d_{ave} (channels 1, 3, 4, and 7 of the TM and four texture features: information measure of correlation, difference variance, sum average and variance, and the elevation feature of the DTM) as input to a classifier based on the BRME. To this end, an MLP with the same architecture as the one used in the previous experiment was employed. In addition, the same learning procedure, the same initial weights, and the same convergence criterion were adopted. The posterior probabilities estimated by this MLP were utilized to perform classification following (1). A global cost equal to 8539 was obtained for the test set, with an overall classification error equal to 15.0%.

A comparison of the total costs obtained in the aforementioned experiments points out that the proposed approach made it possible to sharply decrease the cost involved by classification errors (i.e., about 25.8% lower than the cost value obtained by the approach based on the BRME). On the other hand, it provided a slightly higher overall classification error (i.e., 3.8% higher) than the one provided by the approach based on the BRME.

To better analyze the obtained results, the error matrices achieved by the approaches based on the BRMC and the BRME are shown in Tables III and IV, respectively. These matrices were computed by comparing, for each pixel in the test set, the classification maps provided by the two classification approaches with the ground truth. The terms on the diagonals of the matrices give correctly recognized classes, while the other terms give the errors incurred on the pairs of classes. The class-by-class errors are given in the last column of the matrices. By comparing the errors given in the two matrices, one can deduce that, although the approach based on the BRMC slightly increased the overall classification error, it allowed a sharp reduction in the errors on the classes, which, if confused, would have involved high costs. In greater detail, for example, the confusion between the class of a mixed coppice-pine forest, which is a critical class (7 being its risk value), and the classes of a pasture-land, a pasture-land with trees, and an urban and agricultural land (which are all associated with a risk level equal to 1) was significantly reduced. This is a consequence of the high costs associated with such kinds of errors. On the other hand, the confusion between the class of a mixed coppice-pine forest and the class of a chestnut forest (which is not a critical confusion, since the chestnut forest class is associated with a risk value, i.e., (8), that is very close to the risk level of a mixed coppice-pine forest) slightly increased.

Another interesting example to understand the peculiarities of the proposed approach concerns the class of a dense high maquis. By using the approach based on the BRMC, the errors incurred on this class increased, as compared to the ones made by the approach based on the BRME. This is due to a sharp increase in the confusion between the class of a dense high maquis and the class of a mixed coppice-pine forest. However, these kinds of errors are not critical in the considered application, because a mixed coppice-pine forest is associated with a risk level slightly higher than the one of a dense high maquis (i.e., 7 versus 5).

Both the proposed approach and the approach based on the BRME were then applied to the whole images to derive classification maps and hence, the related risk maps. The obtained risk maps are shown in Figs. 6 and 7. Comparisons between the two risk maps highlight that the proposed approach involves a larger number of pixels belonging to areas associated with high risk values, as compared to the pixels obtained by the approach based on the BRME. This behavior is a consequence of the high cost assigned to the confusion of classes characterized by high risk values with classes characterized by low risk values. In fact, this "favors" classes associated with high risk values in the classification process.

Finally, in order to evaluate separately the effectiveness of the proposed feature-selection criterion function and that of the proposed classification technique, experiments were carried out by giving features selected by d_{cost} as input to the classifier based on the BRME and by giving features selected by d_{ave} as input to the classifier based on the BRMC. The total classification costs and the overall classification errors obtained are summarized in Tables V and VI, respectively. An analysis of these tables points out that, once the classification technique had been chosen, the features selected by d_{cost} made it possible to obtain a lower classification cost than the features selected by d_{ave}. On the other hand, the features selected by d_{ave} resulted in a lower classification error. Once the feature-selection criterion had been fixed, it was possible to observe that the classifier based on the BRMC involved a lower cost than the classifier based on the BRME. On the other hand, the classifier based on the BRME made a lower classification error. These results confirm that, although the proposed feature-selection criterion function and classification technique slightly increased the overall classification error, they allowed a sharp reduction in the total classification cost as compared to the techniques based on the BRME. A further observation derived from these last experiments concerns the interpretation of the feature-selection results. As the only difference between the features selected by the two considered cri-

TABLE III
ERROR MATRIX FOR THE PIXELS OF THE TEST SET FOR THE PROPOSED APPROACH BASED ON THE BRMC

Classified as (BRMC)	True class																Error (%)
	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	ω_{11}	ω_{12}	ω_{13}	ω_{14}	ω_{15}	ω_{16}	
ω_1	89	0	1	0	0	0	1	8	18	21	5	0	0	0	0	0	57.6
ω_2	0	167	30	7	0	2	4	6	2	10	8	2	2	34	2	20	22.0
ω_3	5	11	183	12	0	4	3	3	1	0	1	12	8	5	2	1	28.5
ω_4	0	0	17	425	1	18	0	1	1	0	0	2	41	1	0	1	16.3
ω_5	0	0	0	0	177	0	0	0	0	0	0	2	2	13	4	6	9.2
ω_6	0	0	1	7	8	229	0	0	0	0	0	4	1	0	6	0	12.9
ω_7	1	4	9	0	0	0	72	1	1	19	17	0	0	2	0	1	18.2
ω_8	111	13	1	0	0	0	0	1409	12	126	13	0	0	31	16	43	2.4
ω_9	3	0	0	4	2	1	2	1	196	9	32	6	34	3	1	0	20
ω_{10}	0	0	1	0	0	0	4	1	2	36	15	0	0	0	0	0	84.2
ω_{11}	1	2	4	12	1	0	2	0	10	7	83	5	1	5	1	2	52.8
ω_{12}	0	4	9	15	0	4	0	1	0	0	2	40	37	1	1	0	54.5
ω_{13}	0	0	0	26	0	4	0	0	2	0	0	13	676	0	0	0	15.7
ω_{14}	0	7	0	0	0	0	0	6	0	0	0	2	0	1174	22	13	9.0
ω_{15}	0	4	0	0	5	1	0	1	0	0	0	0	0	19	257	12	21.2
ω_{16}	0	2	0	0	1	0	0	5	0	0	0	0	0	2	14	35	73.9
Overall																	18.8

TABLE IV
ERROR MATRIX FOR THE PIXELS OF THE TEST SET FOR THE CLASSICAL APPROACH BASED ON THE BRME

Classified as (BRME)	True class																Error (%)
	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	ω_{11}	ω_{12}	ω_{13}	ω_{14}	ω_{15}	ω_{16}	
ω_1	80	0	0	0	0	0	0	0	10	4	2	0	0	0	0	0	61.9
ω_2	0	152	21	3	0	2	12	0	2	5	3	4	2	8	0	4	29.0
ω_3	6	22	193	3	0	3	8	0	1	0	2	8	8	0	0	0	24.6
ω_4	0	0	28	450	0	10	0	0	2	0	0	4	14	0	0	0	11.4
ω_5	0	0	0	0	180	0	0	0	0	0	0	1	1	8	2	2	7.7
ω_6	0	0	3	9	8	237	0	0	1	0	1	6	1	0	2	0	9.9
ω_7	0	1	6	0	0	0	58	0	0	3	1	0	0	0	0	0	34.1
ω_8	106	2	0	0	0	0	0	1378	7	78	0	0	0	15	6	24	4.5
ω_9	8	0	0	0	0	0	1	2	178	8	12	5	7	1	1	0	27.3
ω_{10}	4	2	0	0	0	0	4	12	2	97	7	0	0	0	0	0	57.5
ω_{11}	6	1	4	5	0	0	5	3	25	32	148	4	1	0	0	0	15.9
ω_{12}	0	0	1	3	0	4	0	0	0	0	0	28	14	0	0	0	68.2
ω_{13}	0	0	0	35	0	6	0	0	17	0	0	24	754	1	0	0	6.0
ω_{14}	0	22	0	0	1	0	0	16	0	0	0	2	0	1226	24	22	5.0
ω_{15}	0	4	0	0	5	1	0	5	0	1	0	2	0	24	270	13	17.2
ω_{16}	0	8	0	0	1	0	0	27	0	0	0	0	0	7	21	69	48.5
Overall																	15.0

teria lies in the fact that d_{cost} selected the correlation-texture feature, whereas d_{ave} selected band 7 of the TM, one can conclude that, for the considered data set, the correlation-texture feature allows, on average, a better discrimination of the most critical classes than band 7 of the TM. By contrast, band 7 of TM is more effective than the correlation-texture feature if the objective is to minimize the overall error made by the classifier.

VI. CONCLUSIONS

In this paper, an approach to feature selection and classification of remote-sensing images based on the BRMC has been presented. In particular, a feature-selection criterion function and a classification technique have been proposed that take into account the different cost associated with each kind of error.

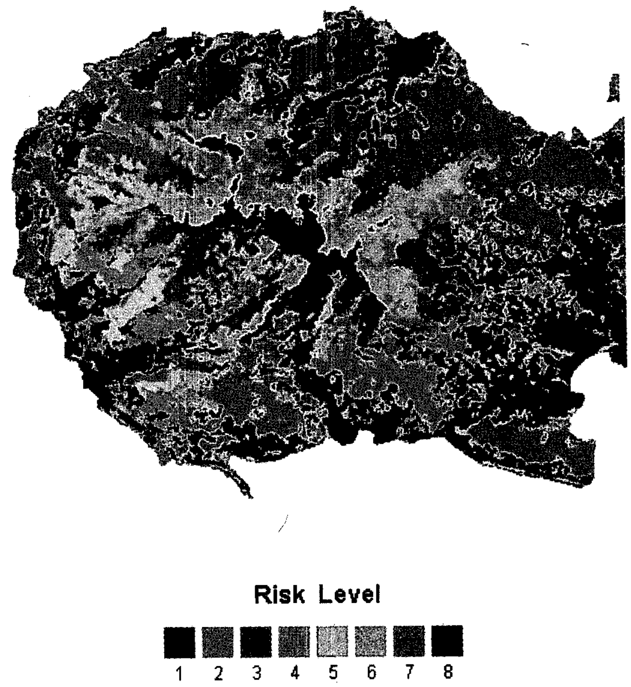

Fig. 6. Risk map obtained in the reported experiments by using the proposed feature-selection criterion function (d_{cost}) and classification technique based on the BRMC. The value 8 is associated with the highest fire probability and the value 1 with the lowest fire probability.

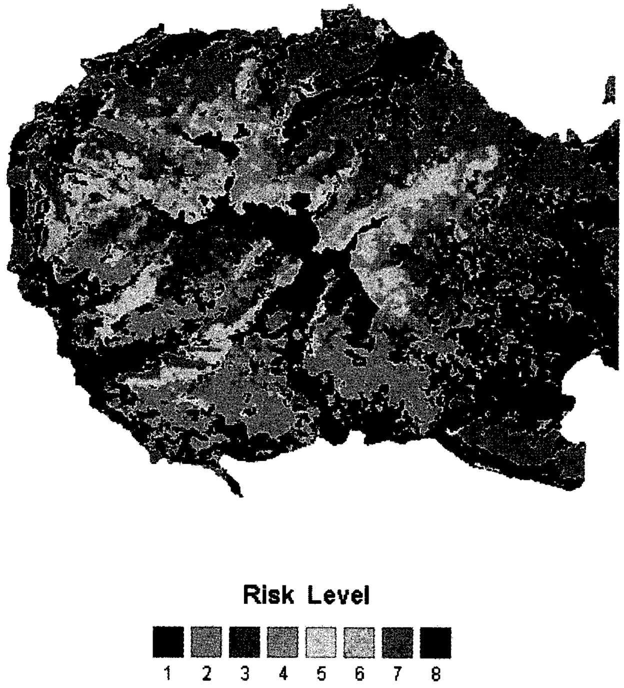

Fig. 7. Risk map obtained in the reported experiments by using the classical feature-selection criterion function (d_{ave}) and classification technique based on the BRME. The value 8 is associated with the highest fire probability and the value 1 with the lowest fire probability.

TABLE V
TOTAL COSTS ASSOCIATED WITH CLASSIFICATION ERRORS MADE ON THE TEST SET IN ALL THE DESCRIBED EXPERIMENTS

Feature-selection criterion	Classification rule	
	BRMC	BRME
Proposed criterion function (d_{cost})	6335	8118
Classical criterion function (d_{ave})	7069	8539

TABLE VI
OVERALL CLASSIFICATION ERRORS MADE ON THE TEST SET IN ALL THE DESCRIBED EXPERIMENTS

Feature-selection criterion	Classification rule	
	BRMC	BRME
Proposed criterion function (d_{cost})	18.8%	15.4%
Classical criterion function (d_{ave})	18.7%	15.0%

These methods can be advantageously utilized to obtain land-cover maps suited to being used in the cases in which different kinds of errors involve consequences of different gravity for the application considered.

The proposed approach was tested to solve the problem of producing a land-cover map suited to being used to derive a risk map related to forest fires. The obtained results confirm that the proposed feature-selection criterion function and classification technique, as compared with methods based on the BRME, provided classification maps in which the overall classification error slightly increased, but the total classification cost was sharply reduced.

In the experiments described in this paper, the proposed feature-selection criterion function was applied by using the Jeffreys–Matusita distance between each pair of classes. However, it can be employed by adopting different pairwise separability indexes (e.g., divergence, transformed divergence, Bhattacharyya distance [17]–[19]). Concerning the proposed classification technique, in this paper, the use of an MLP neural network to estimate posterior class probabilities has been suggested. By using such a network, one can perform a nonparametric estimation of posterior probabilities so that it is possible to process multisensor and multisource remote-sensing data. However, one can also utilize other parametric or nonparametric techniques to estimate these probabilities (e.g., the k-nearest neighbor technique [23]).

Besides describing the proposed feature-selection criterion function and classification technique, one of the purposes of this paper has been to point out to the remote-sensing community that, in some applications, it may be more advantageous to use an approach based on the BRMC than an approach based on the widely used BRME. In comparison with the BRME, the BRMC yields less accurate land-cover maps, but it allows decisions that minimize the total cost of errors. Therefore, the choice of one of the two rules should be carefully made on the basis of the specific problem faced and of end-user requirements.

In order to gain a deeper understanding of the differences between the two considered approaches, it should be stressed that the use of the BRMC is equivalent to modifying the *a priori* probabilities of classes in the BRME approach. In particular, the BRMC weights the conditional-density functions of classes by using differential costs in addition to *a priori* probabilities [see (7) and (8)].

As a final remark, is worth noting that the choice of costs in the C matrix is the most critical step in using the approach based on the BRMC, because no general procedures for an efficient selection of costs can be devised. In fact, the values of costs and the proportions among them depend on the specific application considered.

ACKNOWLEDGMENT

The author wishes to thank Dr. C. Conese and Dr. F. Maselli Institute of Agrometeorology and Environmental Analysis for Agriculture-National Research Council, Florence, Italy (IATA-CNR) for providing the ground truth and the digital terrain model of the considered area and for the assistance in defining the risk value associated with each land-cover class.

REFERENCES

[1] J. A. Richards, *Remote Sensing Digital Image Analysis*, 2nd ed. New York: Springer-Verlag, 1993.

[2] S. B. Serpico, L. Bruzzone, and F. Roli, "An experimental comparison of neural and statistical nonparametric algorithms for supervised classification of remote-sensing images," *Pattern Recognit. Lett.*, vol. 17, pp. 1331–1341, Nov. 1996.

[3] J. D. Paola and R. A. Schowengerdt, "A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification," *IEEE Trans. Geosci. Remote Sensing*, vol. 33, pp. 981–996, July 1995.

[4] J. A. Benediktsson, P. H. Swain, and O. K. Ersoy, "Neural network approaches versus statistical methods in classification of multisource remote-sensing data," *IEEE Trans. Geosci. Remote Sensing*, vol. 28, pp. 540–552, July 1990.

[5] L. Bruzzone and D. F. Prieto, "A technique for the selection of kernel-function parameters in RBF neural networks for classification of remote-sensing images," *IEEE Trans. Geosci. Remote Sensing*, vol. 37, pp. 1179–1184, Mar. 1999.

[6] P. Bosdogianni, M. Petrou, and J. Kittler, "Mixed pixel classification with robust statistics," *IEEE Trans. Geosci. Remote Sensing*, vol. 35, pp. 551–559, May 1997.

[7] F. Wang, "Fuzzy supervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sensing*, vol. 28, pp. 194–201, Mar. 1990.

[8] E. Binaghi, P. Madella, M. G. Montesano, and A. Rampini, "Fuzzy contextual classification of multisource remote sensing images," *IEEE Trans. Geosci. Remote Sensing*, vol. 35, pp. 326–340, Mar. 1997.

[9] T. Matsuyama, "Knowledge-based aerial image understanding systems and expert systems for image processing," *IEEE Trans. Geosci. Remote Sensing*, vol. 25, pp. 305–316, May 1987.

[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed, New York: Academic, 1990.

[11] P. C. Smits, S. Dellepiane, and R. A. Schowengerdt, "Quality assessment of image classification algorithms for land-cover mapping: A review and a proposal for a cost-based approach," *Int. J. Remote Sens.*, to be published.

[12] R. S. Lunetta, R. G. Congalton, L. K. Fenstermarker, J. R. Jensen, K. C. Mcgwire, and L. R. Tinney, "Remote-sensing and geographic information system data integration: Error sources and research issues," *Photogramm. Eng. Remote Sens.*, vol. 57, no. 6, pp. 677–687, 1991.

[13] E. Sali and H. Wolfson, "Texture classification in aerial photographs and satellite data," *Int. J. Remote Sens.*, vol. 13, pp. 3395–3408, Dec. 1992.

[14] L. Bruzzone, C. Conese, F. Maselli, and F. Roli, "Multisource classification of complex rural areas by statistical and neural-network approaches," *Photogramm. Eng. Remote Sens.*, vol. 63, pp. 523–533, May 1997.

[15] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Trans. Comput.*, vol. C-26, pp. 917–922, Sept. 1977.

[16] A. Jain and D. Zongker, "Feature selection: Evaluation, application and small sample performances," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 153–158, Feb. 1997.

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

[18] I. L. Thomas, N. P. Ching, V. M. Benning, and J. A. D'Aguanno, "A review of multi-channel indices of class separability," *Int. J. Remote Sens.*, vol. 8, pp. 331–350, 1987.

[19] P. W. Mausel, W. J. Kramber, and J. K. Lee, "Optimum band selection for supervised classification of multispectral data," *Photogramm. Eng. Remote Sens.*, vol. 56, pp. 55–60, Jan. 1990.

[20] L. Bruzzone, F. Roli, and S. B. Serpico, "An extension of the Jeffreys–Matusita distance to multiclass cases for feature selection," *IEEE Trans. Geosci. Remote Sensing*, vol. 33, pp. 1318–1321, Nov. 1995.

[21] T. Kailath, "The divergence and the Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. COM-15, pp. 52–60, Feb. 1967.

[22] L. Bruzzone and S. B. Serpico, "An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images," *IEEE Trans. Geosci. Remote Sensing*, vol. 35, pp. 858–867, July 1997.

[23] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[24] M. Richard and R. Lippmann, "Neural network classifiers estimate Bayesian *a posteriori* probabilities," *Neural Comput.*, vol. 3, no. 4, pp. 461–463, 1991.

[25] H. Gish, "A probabilistic approach to the understanding and training of neural network classifiers," in *Proc. 1990 Int. Conf. Acoustic, Speech, and Signal Processing*, Apr. 3–6, 1990, pp. 1361–1364.

[26] D. R. Hush and B. G. Horne, "Progress in supervised neural networks," *Signal Process. Mag.*, vol. 10, no. 1, pp. 8–39, 1993.

[27] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Reading, MA: Addison Wesley, 1991.

[28] *Fuzzy Logic and Neural Network Handbook*, McGraw-Hill, New York, 1996.

[29] F. Maselli, A. Rodolfi, L. Bottai, S. Romanelli, and C. Conese, "Classification of mediterranean vegetation by TM and ancillary data for the evaluation of fire risk," *Int. J. Remote Sens.*, to be published.

[30] "General Technical Report INT-194," U.S. Dept. Agriculture, Forest Service, Intermountain Research Station, Odgen, UT, 1986.

[31] R. M. Haralick, K. Shanmugan, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, pp. 610–621, Nov. 1973.

Lorenzo Bruzzone (S'95–M'99) received the Laurea (M.S.) degree in electronic engineering and the Ph.D. degree in telecommunications, both from the University of Genoa, Italy, in 1993 and 1998, respectively.

Since June 1998, he has been a Postdoctoral Researcher at the University of Genoa. He is also currently the Scientific Coordinator of the activities on remote sensing image analysis carried out by the Signal Processing and Telecommunications Group, Department of Biophysical and Electronic Engineering, University of Genoa. His main research contributions are in the area of remote sensing image processing and recognition. In particular, his interests include: feature selection, classification, and change detection. He conducts and supervises research on these topics within the framework of several national and international projects. He has been an evaluator of project proposals within the Fifth Framework Programme of the European Commission. He is the author (or co-author) of more than 50 scientific publications and is a referee for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS, *International Journal of Remote Sensing and Signal Processing*, and *Neural Networks*.

Dr. Bruzzone ranked first place in the Student Prize Paper Competition at the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS'98, Seattle, July 1998). He was recognized as the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING Best Reviewer in 1999. He is a member of the International Association for Pattern Recognition (IAPR).