A Multiscale Expectation-Maximization Semisupervised Classifier Suitable for Badly Posed Image Classification

Andrea Baraldi, Lorenzo Bruzzone, *Senior Member, IEEE*, and Palma Blonda, *Member, IEEE*

***Abstract*—This paper deals with the problem of badly posed image classification. Although underestimated in practice, bad-posedness is likely to affect many real-world image classification tasks, where reference samples are difficult to collect (e.g., in remote sensing (RS) image mapping) and/or spatial autocorrelation is relevant. In an image classification context affected by a lack of reference samples, an original inductive learning multiscale image classifier, termed multiscale semisupervised expectation maximization (MSEM), is proposed. The rationale behind MSEM is to combine useful complementary properties of two alternative data mapping procedures recently published outside of image processing literature, namely, the multiscale modified Pappas adaptive clustering (MPAC) algorithm and the sample-based semisupervised expectation maximization (SEM) classifier. To demonstrate its potential utility, MSEM is compared against nonstandard classifiers, such as MPAC, SEM and the single-scale contextual SEM (CSEM) classifier, besides against well-known standard classifiers in two RS image classification problems featuring few reference samples and modestly useful texture information. These experiments yield weak (subjective) but numerous quantitative map quality indexes that are consistent with both theoretical considerations and qualitative evaluations by expert photointerpreters. According to these quantitative results, MSEM is competitive in terms of overall image mapping performance at the cost of a computational overhead three to six times superior to that of its most interesting rival, SEM. More in general, our experiments confirm that, even if they rely on heavy class-conditional normal distribution assumptions that may not be true in many real-world problems (e.g., in highly textured images), semisupervised classifiers based on the iterative expectation maximization Gaussian mixture model solution can be very powerful in practice when: 1) there is a lack of reference samples with respect to the problem/model complexity and 2) texture information is considered negligible (i.e., a piecewise constant image model holds).**

***Index Terms*—Badly posed image classification, data clustering, generalization capability, image mapping, inductive learning, remotely sensed images, semisupervised samples, supervised learning, texture information, unsupervised learning.**

Manuscript received July 1, 2004; revised July 27, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ivan W. Selesnick.

A. Baraldi was with the Istituto di Studi su Sistemi Intelligenti per l'Automazione, Consiglio Nazionale delle Ricerche (ISSIA-CNR), 70126 Bari, Italy. He is now with the European Commission Joint Research Centre, I-21020 Ispra (Varese), Italy (e-mail: baraldi@ba.issia.cnr.it; a.baraldi@isac.cnr.it; andrea.baraldi@jrc.it).

L. Bruzzone is with the Department of Information and Communication Technology, University of Trento, I-38050 Trento, Italy (e-mail: lorenzo.bruzzone@ing.unitn.it).

P. Blonda is with the ISSIA-CNR, 70126 Bari, Italy (e-mail: blonda@ba.issia.cnr.it).

Digital Object Identifier 10.1109/TIP.2006.875220

I. Introduction

RECENT years have seen substantial developments in new approaches to (unsupervised) data clustering and (supervised) data classification in image processing, pattern recognition, data mining, and machine learning literature [1]–[11]. Unfortunately, the impact of these approaches on their potential field of applications (e.g., the development of commercial image processing software toolboxes) has been scanty [12]–[14]. This lack of impact may be due to: 1) (subjective) functional, operational or computational limitations of the proposed approaches and/or 2) the well-known lack of inherent (objective) superiority of any (supervised) predictive learning classifier as well as any (unsupervised) data clustering algorithm.[1] In fact, on the one hand, the well-known subjective nature of the data clustering problem precludes an absolute judgement concerning the relative efficacy of all data clustering systems [15], [16]. On the other hand, in the supervised learning framework, "if the goal is to obtain good generalization performance in predictive learning" (from a finite labeled data set), "there are no context-independent or usage-independent reasons for favoring one learning or classification method over another" [17, p. 454].

Although well studied, one of the most critical issues in real-world applications of predictive learning methods is the so-called *small sample size problem*, also known as the *Hughes phenomenon* or *curse of dimensionality* [2], [15]–[18]. This problem arises when the size of the available set of reference (labeled, supervised) samples is not sufficient to train an inductive learning algorithm (inducer) effectively, thus causing the induced classifier to be affected by poor generalization capability. In the image classification field this problem becomes even more severe as spatial autocorrelation reduces the informativeness of neighboring pixels by violating the assumption of sample independence [19], which may give rise to the so-called *unrepresentative sample problem* [2]. For example, in recent years, when image understanding started encountering tremendous spatial and spectral complexity, such as in second- and third-generation remote sensing (RS) imagery, it became increasingly difficult, expensive, and/or tedious to collect reference samples having statistical properties appropriate for first-generation classifiers (e.g., maximum likelihood classifiers assuming class-specific normal densities) [2], [19], [20].

[1]The goal of *data clustering* (also known as *exploratory data analysis*) is to separate a finite unlabeled data set at hand into a finite and discrete set of "natural," hidden data structures, termed clusters, on the basis of an often subjectively chosen measure of similarity, i.e., a distance measure chosen subjectively based on its ability to create "interesting" clusters [23].

A possible taxonomy of the bad-conditioning of predictive learning problems is the following (adapted from [2], [21]).

- *Ill-posed* (and *very ill-posed*) *predictive learning problems*: where data dimensionality, D, exceeds the total number M of *independent* representative samples and, as a consequence, is much greater than the number of per-class reference samples m_i, $i = 1, \ldots, L$, where L is the number of classes such that $m_1 + \ldots m_L = M$.
- *Poorly posed predictive learning problems*: where data dimensionality, D, is greater than, or comparable to, the number of (independent) per-class representative samples m_i, $i = 1, \ldots, L$, but smaller than the total number of representative samples M.

To mitigate the small sample size problem, two (complementary nonalternative) strategies can be pursued: 1) input space dimensionality reduction (feature extraction/selection, which will not be considered in the rest of this paper) and 2) development of inductive learning classifiers capable of dealing with a lack of reference samples. Within this second scientific framework, one possible solution consists of exploiting *semilabeled samples* (i.e., unlabeled samples after classification[2]) in adapted versions of the well-known iterative expectation-maximization (EM) maximum-likelihood estimator, such as the sample-based (i.e., context-insensitive[3]) Semisupervised EM (SEM) classifier recently proposed outside of image processing literature [2]. SEM, which is theoretically well founded,[4] as well as its heuristic context-sensitive single-scale version (i.e., suitable for dealing with images), hereafter referred to as contextual SEM (CSEM) [22], appear to be particularly interesting owing to their ability to mitigate the small sample size problem. Although they rely on class-specific normal density assumptions that may be not true in real-world image mapping problems (e.g., in highly textured images) and may require supervision to separate multimodal classes into unimodal subclasses [24], SEM and CSEM provide interesting results in the classification of images featuring only slightly useful texture information [2], [22]. More specifically, SEM's and CSEM's parameter estimation strategies combine the small set of labeled data, whose *explicit class labels* feature full weights, with a large set of semilabeled samples (whose *implicit class labels* may be incorrect, refer to footnote 2) provided with reduced weights.

Unfortunately, it is well known that when the normal model distribution estimated by the iterative (sub-optimal) semisupervised mapping algorithms with EM does not match the true underlying distribution, the large amount of unlabeled data may have an adverse effect on classifier performance on labeled samples (i.e., while pursuing a cost function reduction, these algorithms do not guarantee a better error rate for labeled samples in the next iteration) [25].

Moreover, in badly posed classification problems featuring a very high-input space dimensionality D (ranging from ten up to a few hundreds [21], e.g., in the case of RS hyperspectral data), semilabeled samples alone may not be sufficient to reduce the variance of the covariance matrix estimation process where the number of free parameters increases dramatically (approximately to $0.5 * D^2$). In such cases, recent works recommend class-specific leave-one-out regularized covariance (LOOC) estimators initialized by training samples only. Next, these LOOC estimators are iteratively updated using both semilabeled and training samples until a convergence is reached when a quadratic maximum likelihood (ML) classification output changes very little [21]. Unfortunately, when the number of competing classifiers increases, the computational cost of leave-one-out estimation methods may soon become unaffordable (refer also to Section VI-C). Besides, this robust ML classification approach is not designed to deal specifically with images (i.e., it is noncontextual, neither multi- nor single-scale).

Potentially superior to sample-based SEM and context-sensitive single-scale CSEM in detecting genuine but small image details, an original multiscale heuristic adaptation of the SEM classifier, hereafter identified as multiscale SEM (MSEM), is proposed in this paper. The capability of mitigating the small sample size problem while employing multiscale image analysis mechanisms makes MSEM potentially capable of detecting genuine, but small, structures in piecewise constant or slowly varying color images when little prior knowledge is available. Thus, the potential applicability domain of MSEM is expected to range from, say, mapping RS satellite imagery featuring low (≈1 km) and medium (≈30 m) spatial resolution collected in massive amounts in recent years, to the analysis of biomedical magnetic resonance images (MRIs) [26]. It is noteworthy that to date, either the empirical or the well-founded adaptation of existing sample-based semisupervised classification schemes potentially superior to SEM (such as the recently published cost-effective semisupervised classifier, $\mathrm{CES}^2\mathrm{C}$, conceived as a semisupervised adaptation of the Kernel Fisher's Discriminant (KFD) [25]) to a multiscale image analysis framework appears as an open problem of difficult solution. For example, the sample-based $\mathrm{CES}^2\mathrm{C}$ classifier employs three system parameters (namely, the single-scale spread σ of Gaussian kernels in measurement space, a regularization term γ and a weighting coefficient C in the two-term cost function to be user-defined or estimated by cross-validation over the set of labeled training samples) whose adaptation to a multiscale image classification optimization framework seems quite difficult to conceive.

The rest of this paper is organized as follows. Some notation is introduced in Section II. To provide the paper with tutorial value, context-sensitive image mapping algorithms are surveyed in Section III. Section IV briefly reviews two existing data mapping systems, namely, SEM and the modified Pappas adaptive clustering (MPAC) algorithm [4], [27], [28], both related to MSEM. In Section V, MSEM is proposed as the original contribution of this work. Section VI provides a comparison of MSEM against nonstandard (namely, MPAC, SEM, and CSEM), as well as standard, data mapping approaches in two

[2]Thus, semilabeled samples are as many as the unlabeled samples and available at no extra classification cost. Their *implicit class label*, provided by the classifier, may be incorrect. On the contrary, *explicit class labels* of reference samples, provided by an external *supervisor* or *oracle*, are assumed to be (hard, crisply) correct [2], [18], [22].

[3]Context-sensitive data mapping algorithms, either single- or multiscale, are specifically developed for 2-(spatial) dimensional image mapping tasks, whereas sample-based data mapping algorithms, employing no contextual information, are applicable to any 1-(spatial) dimensional sequence of input patterns.

[4]According to [22], a (suboptimal) iterative predictive learning classifier is defined as theoretically well-founded if it is guaranteed to reach convergence at a (local) minimum of a known cost function.

badly posed image classification experiments. Conclusions are reported in Section VII.

II. Notation

A general notation (mainly adapted from [2]) is established and employed in the rest of this paper. Let us identify with $z_j \in \Re^D$ the jth unlabeled data vector (sample, digital number, pixel in an image), where D is the input space dimensionality and index $j \in \{1, \ldots, N\}$, where N is the total number of unlabeled samples, such that $z = \{z_1, \ldots, z_N\}$ is the observed unlabeled data set. Processed by a data labeling (classification) system (either supervised or unsupervised), each unlabeled sample z_j, $j \in \{1, \ldots, N\}$, can take one implicit (hard, crisp) label, or *discrete status*, $x_j \in \{1, \ldots, L\}$, where L is the total number of labels (equivalent either to the number of clusters in unsupervised data analysis or to the number of thematic classes in supervised data analysis), such that $x = \{x_1, \ldots, x_N\}$ is an arbitrary labeling of the unlabeled data set z.

When a data mapping system provides an unlabeled sample z_j, $j \in \{1, \ldots, N\}$, with a hard implicit label $i \in \{1, \ldots, L\}$, then the unlabeled sample becomes a *semilabeled sample* (refer to footnote 2), identified as $z_{i,j}$. If n_i represents the cardinality of the set of semilabeled samples provided with implicit label i, then $n_1 + \ldots + n_L = N$.

A labeled data set, selected by an external agent (supervisor, oracle), consists of supervised samples $y_{i,k}$, $i \in \{1, \ldots, L\}$, $k \in \{1, \ldots, m_i\}$, where m_i is the number of labeled samples belonging to class i, assumed to be correct, such that $m_1 + \ldots + m_L = M$, where M is the reference sample set cardinality. In general, inequality $M \ll N$ holds.

III. Previous Works in Context-Sensitive Image Labeling

A possible categorization of context-sensitive image labeling (i.e., image classification, clustering and segmentation[5]) algorithms recently proposed in pattern recognition and image processing literature is the following [28].

- *Per-pixel (i.e., noncontextual) classifiers*, either parametric (e.g., Gaussian maximum likelihood) or nonparametric (e.g., the k-nearest neighbor classification rule [15], [16], [29]), *followed by a post-processing low-pass filtering stage*, capable of regularizing the classification solution (i.e., the salt-and-pepper classification noise effect is reduced), based on some empirical criteria or morphological filtering [9], [30], [31]. Although inadequate in detecting fine image details in many real-world problems, this approach is widely adopted due to its conceptual and computational simplicity.
- *Artificial neural networks that employ sliding windows or multiresolution banks of filters in the image domain* (see, for example, [32]–[36]). Neural networks are inductive learning systems featuring important functional properties. They are: 1) distribution-free, i.e., they do not require the data to conform to a statistical distribution known *a priori* and 2) importance-free, i.e., they do not require information on the confidence level of each data source, which is reflected in the weights of the network after training [10], [37], [38]. But the dependence of their results on the shape and size of the processing window (which is usually fixed by the user on an *a priori* basis, i.e., these parameters are neither data-driven nor adaptive) is a well-known problem [33]. To avoid this dependence, a multichannel filtering approach, which is inherently multiresolution, is adopted before classification, e.g., to provide a (nearly) orthogonal decomposition/reconstruction of the raw image [34]–[36]. Proposed applications of multiscale multiorientation filter batteries embrace image analysis and synthesis [39], texture analysis and synthesis [40], [41], texture classification [34], image database retrieval [1] and, therefore, are outside the scope of this paper.
- *Bayesian contextual image labeling systems where maximum a posteriori (MAP) global optimization is pursued by means of Markov random field (MRF)-based local computations* [44]. Because of the local statistical dependence (autocorrelation) of images, there has been an increasing emphasis on the use of statistical techniques based on MRFs capable of modeling image features such as textures, edges and region labels [4], [17], [44], [45]. In MRFs, each pixel—conditioned by its neighbors—is statistically independent of any other pixel. In MAP classification, to enforce spatial continuity in label assignment, an MRF model is imposed on the prior probability term (regularization term), which is combined with a class-conditional probability term. To avoid the computational cost of a simulated annealing technique capable of providing optimal minimization [46], context-sensitive labeling approaches are often combined with the iterative conditional mode (ICM) suboptimal minimization [4], [17], [45]. Based on the assumption that observed pixel gray values are conditionally independent and identically distributed (i.i.d), given their (unknown) labels, the posterior joint probability, $p(x|z)$, can be expressed as [17], [47]

$$p(x|z) \propto p(z_j|x_j)p\left(x_j|\hat{x}_{N_j}\right), \quad j \in \{1, N\}, \quad x_j \in \{1, L\} \tag{1}$$

where $\hat{x}_{N_j}$ is the scene reconstruction in neighborhood N_j centered on pixel j. Equation (1) shows that (image-wide) posterior probability $p(x|z)$ never decreases at any jth pixel-based maximization step. In particular, suboptimal convergence to a local maximum of $p(x|z)$ is guaranteed

[5]Segmentation is the partititioning of image data into nonintersecting areas of connected pixels such that: 1) each region (segment, object) is homogeneous in terms either of color, texture or shape information and 2) the union of no two adjacent regions is homogeneous. Thus, segmentation algorithms exploit simultaneously the pixels' value and position information. Each segment is given a unique digital number (DN) value (per-segment identifier) in the segmented output map [42]. From image processing literature, it is well known that the segmentation problem is ill-posed, i.e., it has a subjective nature [43], just like exploratory data analysis (clustering) or predictive learning from a finite labeled data set. To stress the difference between data clustering and image segmentation, it is worth mentioning that the same segmented map may be generated from different cluster maps. Since the goal of image segmentation is to partition the image data at hand rather than provide an accurate characterization of unobserved (future) samples generated from the same probability distribution, the task of segmentation (like that of clustering) falls well outside the predictive learning framework. It is important to stress that the rest of this paper deals with no segmentation approach, but with image clustering and classification algorithms exclusively.

if, for each pixel $j \in \{1,\ldots,N\}$, ICM estimates label x_j that maximizes the right side of (1), where only the class-conditional probability $p(z_j|x_j)$ and labels of the pixel neighbors $\hat{x}_{N_j}$ are required. An ICM optimization procedure that scans the image iteratively relates the "hat" in (1) to the use of estimated label assignments from the previous iteration in the current iteration, such that batch label updating can be enforced at the end of each raster scan. In other words, ICM alternates between pixel labeling, based on (1), and category-specific model parameter estimation [17], [47]. An interesting ICM algorithm is the MRF-based contextual version of the SEM classifier (CSEM, see Sections I and IV-B), capable of mitigating the small training sample size problem [22]. In [45], different texture regions are modeled by Gauss-MRFs (GMRFs) whose parameters are approximated at various resolutions although the Markov property is lost under such resolution transformation. Unfortunately, GMRFs are good at describing a variety of smooth textures, but perform poorly when sharp edges or small isolated features are to be preserved [48]. In [4], after speculating that an MRF model of the labeling process is not very useful unless it is combined with a good model for class-conditional densities, Pappas presents an ICM-based context-sensitive algorithm for quantization error minimization, hereafter referred to as the Pappas adaptive clustering (PAC) algorithm. PAC adopts a context-sensitive single-scale class-conditional intensity average estimate based on a slowly varying or piecewise constant image intensity model. Unfortunately, PAC tends to remove genuine but small image details. To improve PAC in terms of genuine but small region detection capability, the MPAC algorithm adopts a multiscale class-conditional intensity average estimate [17].

IV. Review of the MPAC and SEM Data Mapping Algorithms

To make this paper self-contained (refer to Section I), this section briefly sketches existing sample-based SEM and multiscale MPAC data mapping approaches to highlight their legacy to MSEM (to be presented further in Section V).

A. The MPAC Contextual Clustering Algorithm

To overcome PAC's well-known limitation, which is that of removing genuine, but small, image regions (refer to Section III, last paragraph), MPAC pursues a multiscale adaptation of the single-scale category-specific intensity average estimation strategy proposed by PAC (see Fig. 1), where texture (correlation) information is assumed to be negligible. In other words, MPAC (like PAC) is exclusively applicable to piecewise constant or slowly varying color images, *possibly* affected by an additive white Gaussian noise field independent of the scene [4]. Let us consider pixel $j \in \{1,\ldots,N\}$ and identify with symbol $\hat{\mu}_{Wj,s}(x_j)$ the slowly varying intensity function estimated as the average of the gray levels of pixels that belong to region type $x_j \in \{1,L\}$ and fall inside an adaptive (local) window $W_{j,s} \subset W$, centered on pixel j at spatial scale $s \in \{1,\ldots,S\}$, where the nonadaptive window W, representing the global scale of analysis, may overlap with the whole image z. The width of window $W_{j,s}$, $\forall j \in \{1,N\}$, $s = 1,\ldots,S$, is identified with symbol WW_s, such that the window width WW_s increases with spatial scale $s \in \{1,S\}$, i.e., $WW_1 < WW_2 < \ldots < WW_S < WW_W$. Symbol β identifies a user-defined (free) parameter (MRF two-point clique potential) enforcing spatial continuity in pixel labeling, such that $\beta \propto \sigma^2$, where σ is the additive white Gaussian noise standard deviation [4], [17]. Given these symbols, the MPAC cost function to be minimized becomes

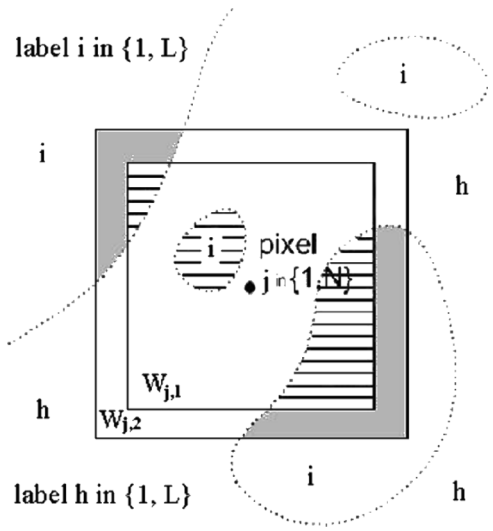


Fig. 1. Multiscale MPAC intensity averages estimation strategy, whose soft-competitive adaptation is employed by the novel MSEM technique (see Section V). In this example, category-conditional intensity averages are extracted from adaptive neighbors within windows Wj,s at spatial scale $s \in \{1,2\}$, centered on pixel $j \in \{1,\ldots,N\}$, where window width WW2>WW1; refers to the textured area with horizontal lines at spatial scale 1, to be added to the gray area at spatial scale 2, for label types i and $h \in \{1,\ldots,L\}$, with $i \neq h$.

$$\hat{x}_j = \arg \min_{x_j \in \{1,L\}} \{\Delta(x_j) + \beta \cdot \hat{v}_j(x_j)\}, \quad j = 1,\ldots,N \tag{2}$$

where the second-order MRF-based cross-aura measure $\hat{v}_j(x_j) \in \{0,8\}$ computes the number of 8-adjacency neighbors of pixel j whose label is different from pixel status x_j, whereas

$$\begin{cases} \Delta(x_j) = \min\left\{[z_j - \hat{\mu}_{Wj,1}(x_j)]^2, \ldots, \right. \\ \quad \left. [z_j - \hat{\mu}_{j,S}(x_j)]^2, [z_j - \hat{\mu}_W(x_j)]^2\right\}, \\ \text{if local statistic } \hat{\mu}_{Wj,s}(x_j) \text{ exists and} \\ \quad \text{is considered reliable, } \forall\, s \in \{1,S\} \\ \Delta(x_j) = [z_j - \hat{\mu}_W(x_j)]^2, \\ \text{if no local statistic } \hat{\mu}_{Wj,s}(x_j) \text{ does exist and} \\ \quad \text{is considered reliable, } \forall\, s \in \{1,S\} \end{cases} \tag{3}$$

where any local estimate $\hat{\mu}_{Wj,s}(x_j)$ is (empirically) considered unreliable if the number of pixels of type x_j, within window $W_{j,s}$ is less than window width WW_s. In cascade to the hard label assignment rules (2) and (3), the second stage of MPAC, which performs multiscale estimation of category-conditional intensity averages $\hat{\mu}_{Wj,s}(x_j)$, $j = 1,\ldots,N$, $x_j = 1,\ldots,L$, $s = 1,\ldots,S$, is shown in Fig. 1. According to (2) and (3) and

to the multiscale intensity averages estimation stage shown in Fig. 1, MPAC may tolerate the same label type to feature different intensity averages in parts of the image separated in space by more than $WW_1/2$, i.e., half of the width of the investigation window that works at the finest resolution (i.e., at spatial scale $s = 1$). While this property guarantees that MPAC is less sensitive to changes in the user-defined number of input clusters than traditional noncontextual (i.e., sample-based) clustering algorithms, such as the hard c-means (HCM) vector quantizer [4], when MPAC reaches convergence, separate image areas featuring different spectral responses may be associated with the same label type. This may lead MPAC to detect artifacts, i.e., to generate an oversegmented output map [28].

B. The SEM Classifier

To mitigate the small training sample size problem, SEM relies on an original EM-based iterative algorithm for maximum likelihood (ML) estimation of Gaussian mixture parameters, where (few) labeled samples are given full weight, and (many) semilabeled samples are given partial weight (refer to Sections I and II). SEM is theoretically well founded (refer to footnote 4) [2]. The description of SEM is summarized as follows.

$0)_{\mathrm{SEM}}$. Initialize Gaussian mixture parameters ϕ_i^c, $i \in \{1, L\}$. Set $c = 0$.

$1)_{\mathrm{SEM}}$. E-Step: compute class-conditional probabilities, $f(z_j|\phi_{i,J}^c)$, $i \in \{1, L\}$, $j \in \{1, N\}$, and weighting factors $w_{ci,j}$, equivalent to relative memberships $r_{i,j}^c$ (which employ neither global nor local priors), $i \in \{1, L\}$, $j \in \{1, N\}$

$$r_{i,j}^c = w_{i,j}^c = \frac{f\left(z_j|\phi_i^c\right)}{\sum_{k=1}^{L} f\left(z_j|\phi_k^c\right)}, \quad r_{i,j}^c \in [0,1]$$

$$\sum_{i=1}^{L} r_{i,j}^c = 1, \quad i \in \{1, L\}, \quad j \in \{1, N\}. \tag{4}$$

$2)_{\mathrm{SEM}}$. Hard sample labeling based on the ML assignment rule

$$\begin{aligned} &x_j \in \text{class label } i, \text{ i.e., (unlabeled)} z_j \Rightarrow \text{(semilabeled)} \\ &\quad z_{i,j}, \text{ to be employed in (7) and (8)} \\ &if\ i = \underset{1 \le h \le L}{\arg\max} \left\{ r_{h,j}^c \right\}, \quad j \in \{1, N\} \end{aligned} \tag{5}$$

$3)_{\mathrm{SEM}}$. M-Step: maximize the mixed log-likelihood

$$\phi_i^+ = \left(\mu_i^+, \Sigma_i^+\right) \in \underset{\phi_i \in \Omega}{\arg\max} \left\{ \sum_{k=1}^{m_i} ln\left(f\left(y_{i,k}|\phi_i^c\right)\right) + \sum_{j=1}^{n_i} w_{i,j}^c ln\left(f\left(z_{i,j}|\phi_i^c\right)\right) \right\}. \tag{6}$$

Thus, the Gaussian mixture parameter update equations become

$$\begin{aligned} \mu_i^+ &= \frac{\sum_{k=1}^{m_i} y_{i,k} + \sum_{j=1}^{n_i} w_{i,j}^c z_{i,j}}{m_i + \sum_{j=1}^{n_i} w_{i,j}^c} \\ &= \frac{m_i}{m_i + \sum_{j=1}^{n_i} w_{i,j}^c} \frac{\sum_{k=1}^{m_i} y_{i,k}}{m_i} + \frac{\sum_{j=1}^{n_i} w_{i,j}^c}{m_i + \sum_{j=1}^{n_i} w_{i,j}^c} \\ &\quad \times \frac{\sum_{j=1}^{n_i} w_{i,j}^c z_{i,j}}{\sum_{j=1}^{n_i} w_{i,j}^c}, \quad i \in \{1, L\} \end{aligned} \tag{7}$$

and see also (8), shown at the bottom of the page.

$4)_{\mathrm{SEM}}$ Check for convergence. If convergence is reached, stop. Otherwise: $c = c + 1$, and goto Step $1)_{\mathrm{SEM}}$.

Limitations of SEM, shared with its heuristic context-sensitive single-scale adaptation, CSEM [22], are pointed out in Section I.

V. Novel MSEM Algorithm for Image Clustering and Classification

Alternative to the existing context-sensitive single-scale empirical CSEM classifier specifically developed for image mapping applications, MSEM is proposed as an original heuristic MPAC-based (i.e., adjustive multiscale) adaptation of the context-insensitive well-founded SEM classifier. On the one hand, as inherited from MPAC, multiscale parameter estimation capabilities should allow MSEM to perform better than SEM in detecting genuine but small image details while avoiding the oversegmentation phenomena that occasionally

$$\begin{aligned} \sum_i^+ &= \frac{\sum_{k=1}^{m_i} \left(y_{i,k} - \mu_i^+\right)\left(y_{i,k} - \mu_i^+\right)^T + \sum_{j=1}^{n_i} w_{i,j}^c \left(z_{i,j} - \mu_i^+\right)\left(z_{i,j} - \mu_i^+\right)^T}{m_i + \sum_{j=1}^{n_i} w_{i,j}^c} \\ &= \frac{m_i}{m_i + \sum_{j=1}^{n_i} w_{i,j}^c} \frac{\sum_{k=1}^{m_i} \left(y_{i,k} - \mu_i^+\right)\left(y_{i,k} - \mu_i^+\right)^T}{m_i} + \frac{\sum_{j=1}^{n_i} w_{i,j}^c}{m_i + \sum_{j=1}^{n_i} w_{i,j}^c} \frac{\sum_{j=1}^{n_i} w_{i,j}^c \left(z_{i,j} - \mu_i^+\right)\left(z_{i,j} - \mu_i^+\right)^T}{\sum_{j=1}^{n_i} w_{i,j}^c}, \quad i \in \{1, L\}. \end{aligned} \tag{8}$$

affect MPAC (see Section IV-A). On the other hand, MSEM should inherit from SEM: 1) an applicability domain extended to supervised as well as unsupervised classification problems; 2) the capability of mitigating the small sample size problem by exploiting supervised as well as semisupervised samples; and 3) an EM-driven guarantee to reach convergence. By employing no (wishful *ad-hoc*) user-defined parameter different from MPACs, MSEM exhibits a data-driven parameter adaptivity (ease of use) equivalent to that of MPAC and superior to those of several competing context-sensitive classifiers, such as CSEM and ICM-MAP-MRF (see Section VI-D), which depend heavily on an empirical smoothing coefficient required to enforce label continuity.

In mathematical terms, MSEM is conceived as an empirical combination of (4), (7), and (8) of SEM (see Section IV-B) with a soft-competitive version of the multiscale objective function adopted by MPAC [refer to (2) and (3)]. Let us identify with $\mathrm{Neigh}_{j,s}$ the adaptive window chosen at spatial scale $s \in \{1, S\}$ and centered on the jth pixel, $j \in \{1, \ldots, N\}$ (refer to Fig. 1). In place of symbol $W_{j,s}$ adopted by the hard-competitive MPAC algorithm (see Section IV-A), symbol $\mathrm{Neigh}_{j,s}$ is adopted herein to identify a neighborhood centered on pixel j, at spatial scale s, featuring soft (relative), rather than hard (e.g., binary), membership values. The proposed MSEM algorithm consists of the following blocks.

$0)_{\mathrm{MSEM}}$. Initialize Gaussian mixture parameters ϕ_i^c, $i \in \{1, L\}$. Set $c = 0$.

$1)_{\mathrm{MSEM}}$. E-Step: compute class-conditional probabilities, $f(z_j|\phi_i^c)$, $i \in \{1, L\}$, $j \in \{1, N\}$, and weighting factors $w_{i,j}^c$, equivalent to relative memberships $r_{i,j}^c$ (which employ neither global nor local priors), $i \in \{1, L\}$, $j \in \{1, N\}$, computed via (4) of SEM.

$2)_{\mathrm{MSEM}}$. Per-pixel hard labeling based on an objective function maximization where multiscale, class-specific intensity averages are weighted by their reliability factors. Compute

$$\text{absolute membership } \eta_{i,j,s}^c = \frac{1}{\mathrm{EuclDis}_{i,j,s}^c + 1} \in [0,1] \quad i \in \{1, L\}, j \in \{1, N\}, s \in \{1, S\} \tag{9}$$

where the absolute membership function (9) is widely employed in pattern recognition literature, e.g., [49]–[52], such that

$$\mathrm{EuclDis}_{i,j,s}^c = \left\| z_j - \mu_{i,j,s}^c \right\|^2 \tag{10}$$

where symbol $\|.\|$ identifies the Euclidean distance (L_2-norm), and

$$\text{intensity average estimate } \mu_{i,j,s}^c = \frac{\sum\limits_{p \in \mathrm{Neigh}_{j,s}} r_{i,p}^c x_p}{\mathrm{SumR}_{i,j,s}^c} \tag{11}$$

where $\mathrm{Neigh}_{j,s}$ is the neighborhood centered on pixel j at spatial scale s, variable $r_{i,p}^c$ identifies a relative membership value computed according to (4), while the normalization factor $\mathrm{SumR}_{i,j,s}^C$ is computed as

$$\begin{aligned} \mathrm{SumR}_{i,j,s}^c &= \sum_{p \in \mathrm{Neigh}_{j,s}} r_{i,p}^c, \text{ such that } \sum_{i=1}^{L} \mathrm{SumR}_{i,j,s}^c \\ &= |\mathrm{Neigh}_{j,s}|, j \in \{1, N\}, s \in \{1, S\}, \text{ holds true} \end{aligned} \tag{12}$$

where $|\mathrm{Neigh}_{j,s}|$ is the cardinality of neighborhood $\mathrm{Neigh}_{j,s}$. Moreover, reliability factors of multiscale class-specific intensity averages are computed as

$$\begin{aligned} &\text{reliability factor } RF_{i,j,s}^c = \frac{1}{|\mathrm{Neigh}_{j,s}|} \mathrm{SumR}_{i,j,s}^c \\ &\text{such that } RF_{i,j,s}^c \in [0,1], \sum_{i=1}^{L} RF_{i,j,s}^c = 1 \\ &j \in \{1, N\}, s \in \{1, S\}. \end{aligned} \tag{13}$$

Equations (9) and (13) are combined into the MSEM objective function as follows:

$$\begin{aligned} &x_j \in \text{class label } i, \textit{ i.e.}, \text{ (unlabeled) } z_j \Rightarrow \text{(semilabeled)} \\ &z_{i,j}, \text{ to be employed in (7) and (8)} \\ &\text{if } i = \underset{1 \le h \le L}{\arg\max} \left\{ \sum_{s=1}^{S} RF_{h,j,s}^c \cdot \eta_{h,j,s}^c \right\}, \; j \in \{1, N\}. \end{aligned} \tag{14}$$

Thus, the MSEM objective function (14) consists of a soft (weighted) combination of multiscale category-specific intensity average estimates, where weighting coefficients are the estimates' reliability factors. These reliability factors take their inspiration from those adopted in multitemporal/multisource optimization problems, where data sources are weighted depending on their different discrimination ability (e.g., refer to [53]). In the case of the MSEM objective function, the role of reliability factors is to measure the degree of compatibility of class-specific statistics estimated at local spatial scales, inherently prone to the small sample size problem, with class-specific statistics estimated at the global (image-wide) spatial scale. In other words, during pixel labeling, MSEM requires multiscale class-specific intensity averages to be consistent through scale. It is noteworthy that objective function (14) employs absolute, rather than relative, memberships [computed by (4)] to avoid the well-known "probabilistic (relative) membership problem." From fuzzy set theory, it is well known that an outlier tends to have small "possibilistic" (absolute) membership values with respect to all category prototypes (models), while its "probabilistic" (relative) membership values may be high [49]–[52].

$3)_{\mathrm{MSEM}}$. M-Step: update Gaussian mixture parameters according to SEM's (7) and (8).

$4)_{\mathrm{MSEM}}$. Check for convergence. If convergence is reached, stop. Otherwise: $c = c + 1$, and goto Step $1)_{\mathrm{MSEM}}$.

To summarize, MSEM shares with SEM (4), (7) and (8) exclusively. A first competitive advantage of MSEM over MPAC is that objective functions (9)–(14), consisting of a weighted combination of class-specific multiscale intensity average estimates, should avoid the detection of artifacts (this phenomenon affecting MPAC as it requires no consistency between inter-scale category-specific mean intensity estimates, see Section IV-A). A second competitive advantage of MSEM over MPAC is that the former applies to both unsupervised and supervised image labeling tasks, i.e., MSEM can be employed with or without a reference labeled data set [in the latter case, no labeled sample with full weight is passed on to (7) and (8)]. In the case of supervised learning tasks, MSEM is expected to mitigate the small sample size problem, in line with SEM and CSEM, by adopting Gaussian distribution parameter update (7) and (8). Another interesting feature of MSEM is that by combining MPAC's with SEM's learning strategies, it pursues robust statistics estimation at local as well as global (image-wide) spatial scales. In particular: 1) at local scales (see Fig. 1), MSEM employs intensity averages which are less sensitive than variance to the small sample size problem, in line with MPAC and 2) MSEM exploits semilabeled samples to mitigate the small sample size problem in the estimation of Gaussian mixture parameters at the global (image-wide) scale, in line with SEM.

A first disadvantage of MSEM with respect to SEM, CSEM and MPAC is its superior computational load. A second drawback of MSEM is that unlike SEM, it benefits from no rigorous statistical foundation, i.e., it is not well-founded (refer to footnote 4). In fact, per-pixel hard labeling equations, (9)–(14), as well as the Gaussian mixture parameter update rules, (7) and (8), are based on heuristics rather than derived from an objective function minimization, e.g., see (6). Since class-specific normal density assumptions adopted by the iterative EM-based semisupervised mapping algorithms are clearly not flexible enough to capture complex image structures encountered in some real-world images, the application domain of MSEM is the same as CSEMs (consisting of two–dimensional (2-D) images where texture information is negligible), equivalent to a subset of SEMs (which is extended to generic one–dimensional (1-D) sequences of multivariate data samples, see Section IV-B).

VI. Badly Posed Classification Session Design: Test Images, Evaluation Measures and Competing Classifiers

This section mainly focuses on the assessment of competing nonstandard MPAC, SEM, CSEM and MSEM classifiers employed in the badly posed classification of images. To satisfy the experimental session validity criteria proposed in [55] and [56],[6] at least two real and standard image classification problems, a battery of measures of success and at least one existing (well-known) data labeling algorithm ought to be selected for comparison purposes [54].

[6]Quite surprisingly, not only 78% of the articles published in the top neural network journals until 1995 [55], but also a large segment of the papers recently published in pattern-recognition literature, e.g., [25], do not satisfy these rather low experimental standards.

A. Empirical Rules to Avoid the Badly Posed Classification of RS Images

Let us first consider the problem of badly posed classification in RS data understanding. In recent years, enhanced spectral, temporal, and spatial resolutions of RS sensors have increased the number of detectable land cover classes. These developments have dramatically increased the size of the ground-truth regions of interest (ROIs) required to be representative of the true class-conditional distributions. Unfortunately, representative samples are expensive, difficult, and/or tedious to digitize from up-to-date reference data acquired from topographic maps, manually interpreted aerial photographs and/or by ground observations. Thus, in RS data mapping, heuristic rules are traditionally adopted to avoid the reference data sampling scheme affected by bad-posedness.

Training (Learning) Phase:

- To avoid the curse of dimensionality, given the number of spectral bands D, general rules of thumb, irrespective of the classifier's free parameters, require that the minimum number m_i of independent training samples belonging to each class $i = 1, \ldots, L$, where L is the total number of classes, be:
  - $m_i \in \{5 * D, 100 * D\}$ [18], [24], [57]. For example, this rule ensures an adequate estimation of nonsingular/invertible class-specific covariance matrices [18];
  - $30 \leq m_i \leq 50$, so that, according to a special case of the central limit theorem, the distribution of many sample statistics becomes approximately normal [19], [20], [58].
- To avoid poor generalization capability of an induced classifier related to model complexity, the minimum number of independent training samples should be proportional to the number of the learning system's free parameters to be optimized during training. For example, for a two-layer network of threshold units, an approximate worst-case bound on generalization is that correct classification of a fraction $1 - \varepsilon$ of new examples requires a number of training patterns at least equal to $M_{\text{train}} \approx F/\varepsilon$, where F is the total number of the system's free parameters. If $\varepsilon = 0.1$, we need around ten times as many training patterns as there are free parameters in the network [15].
- Representative samples should be capable of representing all possible variations in spectral response in each land cover type of interest.

Testing Phase: When overall classification accuracy, $p_c \in [0, 1]$, with an error tolerance of $\pm\delta$ is given as a project requirement, it is possible to estimate the (unobserved) testing sample set size M_{test} as [59]

$$M_{\text{test}} = \frac{(1.96)^2 \cdot p_c \cdot (1 - p_c)}{\delta^2}, \text{ where } p_c \text{ and } \delta \in [0, 1], p_c \gg \delta. \tag{15}$$

In this context, if a holdout resampling method is adopted for the generalization capability assessment of competing classifiers where 1/3 of the reference data set is employed for testing [60], then the overall reference sample set size becomes $M = M_{\text{train}} + M_{\text{test}} = 3 * M_{\text{test}} = 3 * (15)$.

(a) (b)

Fig. 2. (a) Test case 1. False color composition (B: VisBlue, G: NearIR, R: VisRed) of the SPOT image of Porto Alegre, Brazil, 512 × 512 pixels in size, four-band, 20-m spatial resolution, acquired on Nov. 7, 1987. (b) Test case 2. True color composition (B: VisBlue, G: VisGreen, R: VisRed) of the seven-band Landsat TM image provided by the GRSS Data Fusion Committee, 750 × 1024 pixels in size, 30 m spatial resolution.

B. Test Set of RS Images

According to [54], a test set of RS images, suitable for comparing the performance of algorithms employed in image understanding tasks, should be: 1) as small as possible; 2) consistent with the aim of testing; 3) as realistic as possible; and 4) such that each member of the set reflects a given type of image encountered in practice. In this work, the test set of RS images consists of two real-world satellite images, characterized by different sizes and dimensionalities, fragmentation (i.e., visual complexity, related to the presence of genuine but small image details), and levels of prior knowledge, ranging from ill- to poorly posed (refer to Section I).

The raw image adopted in test case 1 is shown in Fig. 2(a). This is a four-band SPOT image of the city area of Porto Alegre (Brazil), 512 × 512 pixels in size, featuring a spatial resolution of 20 m [17]. The image employed in test case 2 is shown in Fig. 2(b). It is a seven-band Landsat TM image, 750 × 1024 pixels in size, with a spatial resolution of 30 m, depicting a country scene in Flevoland (The Netherlands). This second test image is extracted from the standard grss_dfc_0004 data set provided by the GRSS Data Fusion Committee [61]. In visual terms, the presence of nonstationary image structures, such as step edges and lines, combined with many genuine but small image details, makes the town scene more fragmented than the country scene. Both test images are considered as piecewise constant or slowly varying intensity images featuring little useful texture information, i.e., reference ROIs localized and identified in test cases 1 and 2 correspond to *spectrally, rather than texturally, uniform areas of interest*. Moreover, in both test cases 1 and 2, each reference ROI identifies a distinct surface class of interest (which is a rather common practice in real-world RS applications [6]). Twenty-one ROIs/classes are identified in Fig. 2(a) [see Table I(a)], and 12 ROIs/classes are identified in Fig. 2(b) [see Table I(b)], respectively. It is noteworthy that, according to Sections I and VI-A, if class-specific mean and covariance matrix parameters are to be employed, then test problem 1, where the minimum number of i.i.d. samples per-class would be $10 \times \text{number of free parameters} \approx (10 \times (D + 0.5D^2)) = (10 \times (4 + 8)) = 120$, is rather ill-posed, while test problem 2, where $(10 \times (D + 0.5D^2)) = (10 \times (7 + 24.5)) = 315$, is rather poorly posed (ill-posed if the image autocorrelation, superior to that in test case 1, were considered in violating the hypothesis of i.i.d. samples). The complexity of the classification problem is also increased by the partial overlap between spectral signatures. In test case 1, the minimum Jeffries-Matusita distance between ROI pairs [62], $JM \in [0, 2]$, is that between classes *vegetated area* 1 and *vegetated area* 2, equal to 0.50. In test case 2, the minimum Jeffries-Matusita distance is that between classes *scrub 1* and *scrub 2*, equal to 1.80.

C. Set of Measures of Success

When the small/unrepresentative sample problem occurs (as in test cases 1 and 2), traditional classification error estimation methods soon become unsuitable [15], [16], [29], [57], [59]. In particular, the following.

- The resubstitution method increases its optimistic bias with the small sample size. For example, in [28], the resubstitution error was not in line with qualitative results by expert photointerpreters.
- The holdout method is inefficient in exploiting the available data set for training, i.e., it is unfitted to deal with the small sample size problem.
- The leave-one-out method requires a large computational cost even when the classification problem is badly posed. When the number of competing classifiers increases, the computational cost of the leave-one-out method may soon become unaffordable.

TABLE I
(a) TEST CASE 1: 21 ROIS SELECTED ON THE SPOT IMAGE DEPICTED IN FIG. 2(a). (b) TEST CASE 2: 12 ROIS SELECTED ON THE LANDSAT TM IMAGE DEPICTED IN FIG. 2(b)

ROI	Surface type	No. of pixels (% of the total pixels)
1	*dark artificial target 1*	11 (0.004)
2	*dark artificial target 2*	8 (0.003)
3	*bright artificial target*	9 (0.003)
4	*bridge*	21 (0.008)
5	*road 1*	23 (0.008)
6	*road 2*	35 (0.013)
7	*airport*	105 (0.040)
8	*sea port*	21 (0.008)
9	*building 1*	18 (0.006)
10	*building 2*	17 (0.006)
11	*building 3*	24 (0.008)
12	*vegetated area 1*	21 (0.008)
13	*vegetated area 2*	89 (0.033)
14	*vegetated area 3*	155 (0.059)
15	*vegetated area 4*	124 (0.047)
16	*vegetated area 5*	66 (0.025)
17	*grassland*	26 (0.009)
18	*bare soil*	47 (0.017)
19	*sea water 1*	141 (0.053)
20	*sea water 2*	304 (0.115)
21	*sea water 3*	63 (0.024)
TOTAL		1328

(a)

ROI	Surface type	No. of pixels (% of the total pixels
1	*arable land 1*	689 (0.089)
2	*arable land 2*	571 (0.074)
3	*arable land 3*	787 (0.102)
4	*arable land 4*	238 (0.030)
5	*arable land 5*	1210 (0.157)
6	*vegetated agricultural area 1*	245 (0.031)
7	*vegetated agricultural area 2*	415 (0.053)
8	*vegetated agricultural area 3*	620 (0.080)
9	*vegetated agricultural area 4*	128 (0.016)
10	*scrub 1*	320 (0.041)
11	*scrub 2*	308 (0.040)
12	*marine water*	14900 (1.93)
TOTAL		20431

(b)

- In the n-fold cross validation, the computational load, which increases linearly with the number of competing classifiers by an n factor, may soon become unaffordable.
- The bootstrap method has the largest computational cost, which soon becomes prohibitive with the number of competing classifiers.

Last, but not least, none of these reference data set resampling methods allows estimation of the spatial distribution of classification errors (known as *location accuracy* [58]).

To avoid the aforementioned limitations of traditional resampling techniques, the recently published data-driven map quality assessment (DAMA) strategy can be employed [59]. To mitigate the small and unrepresentative sample problems in estimating and comparing competing classifiers with a mimimum of human intervention, DAMA integrates the available (small) labeled data set, if any, with many semilabeled samples generated *ad hoc* from cluster analysis. As a consequence, the small labeled data set can be efficiently (fully) exploited in training the inducer. In deeper detail, DAMA computes quantitative indexes of labeling and segmentation consistency between the classification map at hand, made from a digital input image, and *reference cluster maps* properly generated from several blocks of the input image that are clustered separately to detect genuine, but small, image details.

In test cases 1 and 2, candidate representative areas are (subjectively) selected as, respectively, three image subsets of Fig. 2(a) (100×300 pixels each), and two image subsets of Fig. 2(b) (400×400 pixels each) (for implementation details, refer to [59]). In combination with the unsupervised DAMA strategy, additional measures of classification success can be conveniently computed in badly posed image classification problems. Since small reference ROIs are available and efficiently exploited for training the inducer, a confusion matrix, computed between the map under investigation and each reference cluster map, allows estimation of the so-called *resubstitution error* (upon the training data set). A necessary (although not sufficient) condition to have good generalization capabilities (i.e., to keep the combination of bias with variance low [15]) is that the optimistically biased resubstitution error be small, i.e., prior knowledge must be successfully passed on to the image mapping system.

A fourth feature that may be considered important in the assessment of competing classifiers is computation time, which affects the application domain of RS image mapping systems [6], [28], and may determine whether an algorithm is capable of enriching a commercial image processing software toolbox, as required by Zamperoni [12].

D. Set of Classifiers to Be Compared

Based on their functional properties, existing classifiers can be partitioned into [15], [29], [50], [57]: 1) context-sensitive (i.e., specialized to deal with images, based on either single-scale or multiscale image analysis mechanisms) and sample- (i.e., pixel-)based; 2) supervised learning, unsupervised learning (clustering methods), and semisupervised learning (see Sections I and IV-B); 3) parametric and nonparametric (also called memory-based [29], whose computational complexity increases with the cardinality of the representative data set); and 4) adaptive and nonadaptive (also called plug-in, where the spectral response for each land cover category of interest is determined off-line, in the training stage, prior to the

TABLE II
TAXONOMY OF THE DATA LABELING ALGORITHMS ADOPTED FOR COMPARISON

Algorithm	Learning mode		Parametric vs Memory-based	Plug-in	Iterative ICM-based vs EM-based	Soft/Hard competitive learning	Context-sensitive
	Sup.	Unsup.					
NP	Y	-	P	Y	N	-	N
ML	Y	-	P	Y	N	-	N
ICM-MAP-MRF	Y	-	P	N	ICM	Hard	Single-scale*
MPAC	Y	-	P	N	ICM	Hard	Multi -scale
SEM	SEM2	SEM1	P	N	EM	Semilabeled	N
CSEM	CSEM2	CSEM1	P	N	EM	Semilabeled	Single-scale*
MSEM	MSEM2	MSEM1	P	N	EM	Semilabeled	Multi-scale

Legend: Y: Yes, N: No. Single-scale$*$: MRF-Based 8-adjacency neighborhood. P: Parametric. M: Memory-based (nonparametric). ICM: Iterative Conditional Mode. EM: Expectation-Maximization.

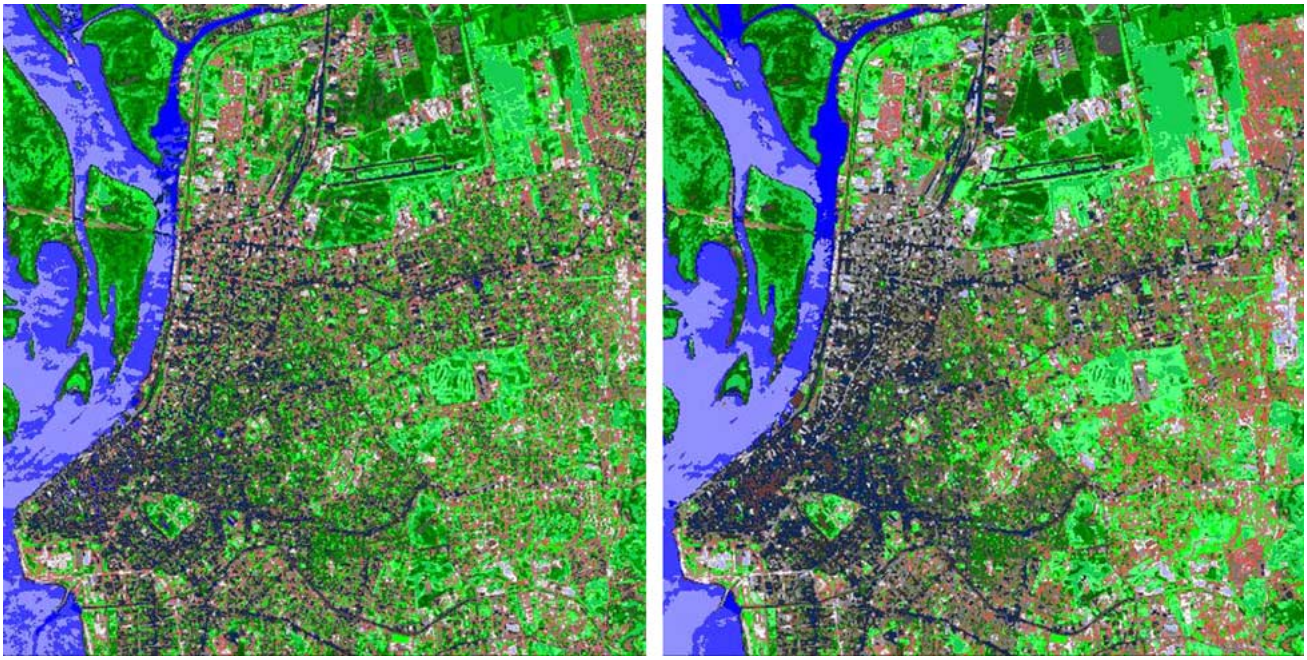
classification stage, by an external analyst who replaces the unknown parameters in the class-specific densities with their estimated values [57]). According to this terminology, labeling systems are either supervised (e.g., multilayer perceptrons [15], radial basis functions [5], probabilistic neural network [63]) or unsupervised (e.g., HCM, enhanced Lloyd–Buzo–Gray [7]), whereas plug-in approaches (like the Bayesian plug-in classifiers, either Bayes-normal-quadratic or Bayes-normal-linear [57]) must be supervised and parametric.

To make the assessment of competing nonstandard mapping systems (namely, MPAC, SEM, CSEM, and MSEM) comprehensive, their comparison must involve well-known (standard) classification algorithms, selected from pattern recognition literature (not necessarily from image processing literature) and/or implemented in commercial image analysis software toolboxes capable of covering a wide range of functional characteristics (see Table II). The standard classifiers selected for comparison purposes are: the iterative conditional mode (ICM)-based maximum a posteriori (MAP)-Markov random field (MRF) classifier [64], the nearest prototype (NP) classifier (also termed mimimum-distance-to-mean classifier [24], implemented in commercial image processing software toolboxes [42]), and the Gaussian maximum likelihood (ML) classifier [15] (it, too, implemented in commercial image processing software toolboxes [42]). Overall, seven different algorithms (implemented in ten versions, see Section VI-D-1) are compared. A rough taxonomy of the compared classifiers is proposed in Table II.

1) Initialization Strategies Exploiting A Small Representative Data Set: In order to guarantee a fair comparison between competing image mapping systems, prior knowledge, having the initial form of reference ROIs, must adapt its maximally informative representation to the learning properties of the system at hand. In our experiments involving many parametric algorithms (either supervised or unsupervised), the number of template vectors (also called reference vectors) is assumed to be coincident with the number of surface types of interest (in a classification framework, these systems are known as one-prototype classifiers [65]). This means that each distribution of class-specific representative samples is assumed to match the class-specific spectral distribution model adopted by the labeling algorithm at hand (e.g., piecewise constant intensity image model for MPAC). For parametric mapping systems, such as ML, ICM-MAP-MRF, SEM, CSEM, and MSEM, this hypothesis implies that there is no need to split multimodal class-specific densities into unimodal normal densities. This assumption, which is typical of first-generation classifiers [20], is likely to hold in low- to medium-spatial resolution satellite images like those adopted in our experiments (see Section VI-B), whereas it would naturally fail in very high spatial resolution (third generation) RS imagery (featuring multimodal distributions in color space and relevant texture information in the 2-D image domain).

Let us model each supervised ROI (corresponding to a spectrally uniform surface area, see Section VI-C) with a Gaussian distribution, parameterized by a (mean vector, covariance matrix) pair, identified as $(\mu_{i,0}, \Sigma_{i,0})$, $i = 1, \ldots, L$. Thus, NP is plugged-in with estimates $\mu_{i,0}$, $i = 1, \ldots, L$. The MPAC clustering algorithm is initialized with mean template vectors $\mu_{i,0}$, $i = 1, \ldots, L$. With regard to parametric plug-in classifiers (e.g., ML) or iterative learning systems that employ class-specific Gaussian distributions (which is the case of ICM-MAP-MRF, SEM, CSEM and MSEM), the empirical rules proposed in Section VI-A recommend that a number of class-specific training samples equal, or possibly superior, to $10 \times \text{number of free parameters} \approx (10 \times (D + 0.5 \cdot D^2))$ be selected to ensure an adequate estimation of a per-class (mean vector, covariance matrix) pair.

To ensure estimation of nonsingular-invertible class-specific covariance matrices from unrepresentative training samples, parametric ML and ICM-MAP-MRF, altogether with a specific implementation of the *partially semisupervised* classifiers SEM, CSEM and MSEM (identified as version SEM2, CSEM2, and MSEM2, where labeled samples with full weight as well as semilabeled samples with partial weight are passed on to their learning phase), as well as their *purely semisupervised* versions (identified as SEM1, CSEM1, and MSEM1, respectively, where semilabeled samples with partial weight exclusively are passed on to their learning phase, i.e., no labeled sample with full weight is employed during training), employ the following semisupervised initialization strategy (at iteration 0). First, supervised ROI-driven mean vector estimates $\mu_{i,0}$, $i = 1, \ldots, L$, are passed on to a nearest-prototype classification step, NP.



(a) (b)

Fig. 3. (a) Test case 1. MPAC clustering of the four-band SPOT image, number of classes $L = 21$, shown in pseudocolors (refer to footnote 7). (b) Test case 1. MSEM1 clustering of the four-band SPOT image, number of classes $L = 21$, shown in pseudocolors (refer to footnote 7).

Next, the hard output map generated from NP provides statistically meaningful image-wide category-specific estimates $(\mu_{i,}, \Sigma_{i,})$, $i = 1, \ldots, L$, which are finally adopted, at iteration 1, by the iterative learning system at hand. It is noteworthy that in the case of poorly posed classification problems, due to the presence of many semilabeled samples (provided with partial weight) and of few labeled samples (provided with full weight, whose contribution to the system's free parameter estimation may become negligible), partially semisupervised implementations SEM2, CSEM2, and MSEM2 are expected to behave somewhat similarly to their purely semisupervised counterparts SEM1, CSEM1, and MSEM1, sharing the same semisupervised initialization strategy.

2) User-Defined Parameter Setting: Context-sensitive multiscale image mapping algorithms (namely, MPAC, MSEM1 and MSEM2), adopt a battery of three local window sizes, equal to 3×3, 7×7, 11×11, to be employed in combination with the global (image-wide) scale, e.g., 512×512 in test case 1 (see Sections IV-A and V). MPAC employs a spatial continuity parameter $\beta = 0$ in (2), to inhibit its MRF-based contextual mechanism, such that its context-sensitivity is exclusively due to multiscale intensity averages estimation. The maximum number of iterations is set equal to ten in the entire set of iterative algorithms. Context-sensitive single-scale MRF-based algorithms (namely, CSEM1, CSEM2, and ICM-MAP-MRF), employ two-point clique potential parameters $\beta = \beta_i = 1.0$, $i = 1, \ldots, L$. It is obvious that optimal smoothing parameters $\beta_h, h \in \{1, L\}$, are both class- and application-dependent. 2To avoid a time-consuming, class-specific, trial-and-error parameter selection strategy that would represent a degree of user's supervision superior to that required by the rest of the algorithms involved in our comparison, we set two-point clique potential parameters $\beta = \beta_i = 1.0$, $i = 1, \ldots, L$, independent of the class. This choice is in line with [64], where $\beta \in [1.0 - 1.6]$, independent of the data set because larger values of β would lead to excessive smoothing of regions.

TABLE III
TEST CASE 1. OVERALL RESUBSTITUTION ACCURACY (SUM OF DIAGONAL ELEMENTS OF THE CONFUSION MATRIX) BETWEEN LABELING RESULTS AND REFERENCE DATA (ROIS). NUMBER OF LABEL TYPES (= number of reference ROIs) = 21

Classifier	Resubstitution overall accuracy %	Rank1
NP	87.8	1
ML	82.1	3
ICM-MAP-MRF	87.4	2
MPAC	65.3	10
SEM1*	66.3	9
SEM2**	69.1	5
CSEM1*	66.9	8
CSEM2**	67.6	7
MSEM1*	68.4	6
MSEM2**	70.9	4

$*$: without supervised (training) samples. $**$: with supervised (training) samples. Rank1 is best when smallest

VII. EXPERIMENTAL RESULTS

Of the systems compared in this experimental session, plug-in NP and ML are expected to perform well in minimizing the resubstitution error (where bias must be low), while parametric iterative (adaptive) labeling algorithms, either unsupervised (MPAC), supervised (ICM-MAP-MRF),

TABLE IV
TEST CASE 1. RESULTS OBTAINED BY THE DAMA STRATEGY FOR MAP QUALITY ASSESSMENT

Classifier	Standard value of the overlapping area (%) on reference cluster map 1 (sample mean = 44.1200, sample std = 9.0314)	Standard value of the overlapping area (%) on reference cluster map 2 (sample mean = 43.7300, sample std = 5.9268)	Standard value of the overlapping area (%) on reference cluster map 3 (sample mean = 38.7600, sample std = 6.6988)	Average of standard mean values 1, 2, and 3	Rank2
NP	-1.6852 (*28.9*)	-1.0343 (37.6)	-1.6809 (27.5)	-1.4668	10
ML	-0.8216 (*36.7*)	-1.6248 (34.1)	-1.0688 (31.6)	-1.1718	8
ICM-MAP-MRF	-1.5967 (*29.7*)	-1.0343 (37.6)	-1.3823 (29.5)	-1.3378	9
MPAC	0.2414 (*46.3*)	1. 8340 (54.6)	0.8718 (44.6)	0.9824	1
SEM1*	0.8393 (51.7)	0.1805 (44.8)	0.2448 (40.4)	0.4216	4
SEM2**	0.5957 (49.5)	0.1974 (44.9)	0.2448 (40.4)	0.3460	5
CSEM1*	0.3300 (47.1)	0.1805 (44.8)	0.3941 (41.4)	0.3015	6
CSEM2**	0.2967 (46.8)	0.0793 (44.2)	0.4240 (41.6)	0.2667	7
MSEM1*	0.8836 (52.1)	0.6192 (47.4)	0.9763 (45.3)	0.8264	3
MSEM2**	0.9168 (52.4)	0.6024 (47.3)	0.9763 (45.3)	0.8318	2

$*$: without supervised (training) samples. $**$: with supervised (training) samples. Rank2 is best when smallest

purely semisupervised (namely, SEM1, CSEM1, and MSEM1), or partially semisupervised (namely, SEM2, CSEM2, and MSEM2), where all semilabeled samples contribute to the adaptation of category-specific template vectors, are expected to improve their generalization ability upon unobserved image areas (when the combination of bias with variance must be kept low) at the cost of a possible increase in their resubstitution error on reference ROIs (due to an increase in bias). Based on model complexity, adaptive MPAC should employ plug-in NP as a reference, whereas purely and partially semisupervised implementations of SEM, CSEM, and MSEM should employ plug-in ML as a reference.

All our experiments are conducted on a workstation SUN Ultra 5 with operating system SunOS 5.6, 64 MB of RAM, and a CPU UltraSPARC-IIi at 270 MHz. No optimization is employed at code compilation.

A. SPOT Image Test Case

As two interesting examples of the mapping results obtained with the proposed parameter setting, Figs. 3(a) and 3(b) show (in pseudocolors[7]) the maps generated with, respectively, clustering algorithms MPAC and MSEM1 (the other output maps are omitted to save presentation space). According to perceptual quality criteria adopted by expert photointerpreters, MPAC and MSEM1 appear to perform better than several other competing systems (in terms of genuine but small image details detection), although their maps look rather different (e.g., in Figs. 3(a) and 3(b), note the different spatial distributions of water types).In the framework of a resubstitution error estimation method, Table III reports the overall accuracy (sum of diagonal elements of the confusion matrix) between labeling results and reference ROIs. Table III shows that in line with theoretical expectations the resubstitution accuracy of some of the parametric iterative labeling algorithms (namely, MPAC, SEM, CSEM, and MSEM), is largely inferior to that of traditional parametric plug-in classifiers (NP and ML). Partially semisupervised classifiers SEM2, CSEM2, and MSEM2 perform better than their purely semisupervised counterparts, in line with theoretical expectations. Although a low resubstitution error is a desirable property, optimistically biased estimates, provided by Table III, are counter-intuitive for expert photointerpreters employing perceptual quality criteria.

To compare the generalization capabilities of predictive learning methods according to the DAMA assessement strategy, Table IV shows the maximum sum (after reshuffling) of diagonal elements of the overlapping area matrix computed between the output map, x, and the multiple cluster maps generated from the raw image, z. In line with qualitative photointerpretation of mapping results, Table IV reveals that labeling fidelities to multiple cluster maps of the MPAC's and MSEM's output maps appear to be superior to those of the other labeling approaches, including NP and ML (as theoretically expected). MSEM performs better than SEM, in line with theoretical expectations, while SEM performs better than CSEM in preserving genuine but small image details, which is theoretically plausible but in contrast with conclusions found in [22]. Overall, in line with theoretical expectations (see Section VI-B), the poor correlation between Rank1 (from resubstitution) and Rank2 (from generalization) reveals the presence of the Hughes phenomenon.

To investigate the spatial fidelity of segmentation results to reference data according to the DAMA strategy, Table V reports the mean of the edge map difference computed as the absolute point-by-point difference between the 4-adjacency edge

[7]Every class index is associated with a pseudocolor chosen to mimic the true color of that surface class (e.g., three shades of blue are adopted to depict labels belonging to classes *sea water 1* to *sea water 3*, etc.), to enhance human interpretability of mapping results.

TABLE V
TEST CASE 1. MEAN AND STANDARD DEVIATION OF THE DIFFERENCE EDGE MAP COMPUTED BETWEEN THE TWO EDGE MAPS MADE FROM x_i^* AND x_i, $i = 1, \ldots, 3$

Classifier	Standard value of the mean (in [0,4]) of edge map difference 1 (sample mean = 0.9450, sample std = 0.1873)	Standard value of the mean (in [0,4]) of edge map difference 2 (sample mean = 0.8750, sample std = 0.1765)	Standard value of the mean (in [0,4]) of edge map difference 3 (sample mean = 1.0520, sample std = 0.2556)	Average of standard mean values 1, 2 and 3	Rank3
NP	1.1477 (1.16)	0.1983 (0.91)	1.0096 (1.31)	0.7852	9
ML	-0.3470 (0.88)	-0.2549 (0.83)	-0.5165 (0.92)	-0.3728	4
ICM-MAP-MRF	2.3754 (1.39)	2.5775 (1.33)	2.3400 (1.65)	2.4310	10
MPAC	-0.8808 (0.78)	-1.1047 (0.68)	-1.0644 (0.78)	-1.0166	1
SEM1*	-0.6139 (0.83)	-0.7081 (0.75)	-0.6731 (0.88)	-0.6650	2
SEM2**	-0.6139 (0.83)	-0.6515 (0.76)	-0.6731 (0.88)	-0.6461	3
CSEM1*	-0.1335 (0.92)	0.0850 (0.89)	0.0704 (1.07)	0.0073	7
CSEM2**	-0.0801 (0.93)	0.1983 (0.91)	0.1096 (1.08)	0.0759	8
MSEM1*	-0.4537 (0.86)	-0.1983 (0.84)	-0.3209 (0.97)	-0.3243	5
MSEM2**	-0.4004 (0.87)	-0.1416 (0.85)	-0.2817 (0.98)	-0.2746	6

$*$: without supervised (training) samples. $**$: with supervised (training) samples. Rank3 is best when smallest

map extracted from the output map and the one extracted from every multiple cluster map. Table V shows that multiscale labeling algorithms (namely, MPAC and MSEM), context-insensitive adaptive SEM and nonadaptive ML are superior to the other algorithms in preserving genuine but small image details, irrespective of their labeling. In particular, MPAC outperforms the other competing systems, whereas SEM performs better than MSEM, which is, in turn, better than CSEM. These spatial fidelity results appear to be in moderate agreement with the labeling accuracy results shown in Table IV, as confirmed by the Spearman correlation coefficient computed between Rank2 and Rank3, equal to 0.6835 (revealing moderate agreement [19]).

Computation time of the competing algorithms is proposed in Table VI, which shows that in this experiment the quality of mapping results appears to be inversely proportional to computation time. In particular, SEM seems to guarantee an interesting compromise between labeling and spatial fidelity of output results to reference data, with computation time.

Overall, these conclusions appear to be consistent with those by expert photointerpreters and in line with the theoretical expectations of the algorithms' potential utility.

B. Landsat Image Test Case

This test image is less fragmented than test case 1. As a consequence, in this experiment, functional benefits deriving from the use of the context-sensitive single-scale ICM-MAP-MRF and CSEM algorithms (provided with an MRF-based mechanism to enforce spatial continuity in pixel labeling) are expected to be superior to those in test case 1. User-defined parameters are the same as those selected in test case 1. As in test case 1, interesting examples of the mapping results obtained with this parameter setting are shown in Figs. 4(a) and (b), where two maps generated by MPAC and MSEM1 respectively, are depicted (in pseudocolors). In test case 2, due to its large fragmentation and to the absence of easy-to-recognize built-up areas, it is rather difficult for expert photointerpreters to determine whether, for example, MPAC [see Fig. 4(a)] performs better than MSEM1 [see Fig. 4(b)].

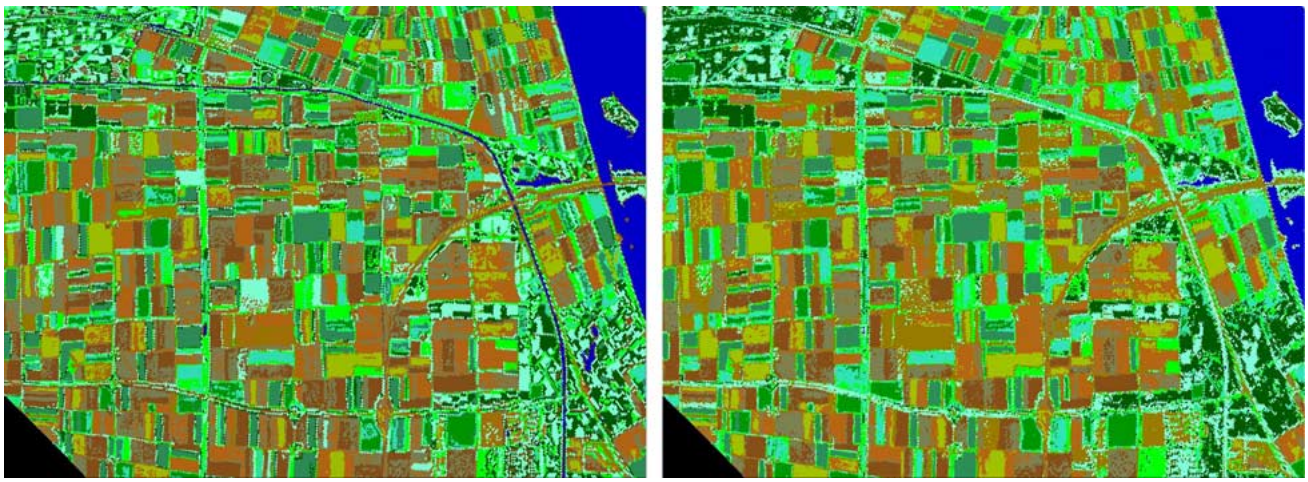
TABLE VI
COMPUTATION TIMES OF THE INDUCTIVE LEARNING ALGORITHMS IN THE SPOT IMAGE TEST CASE

Classifier*	Computation time	Rank4
NP	4s	1
ML	27s	2
ICM-MAP-MRF **	7m 30s	3
MPAC **	28m 30s	8
SEM1, SEM2 **	10m 50s	4.5
CSEM1, CSEM2 **	12m 10s	6.5
MSEM1, MSEM2 **	58m 10s	9.5

$*$: 21 label types, 1328 training pixels (0.5%). $**$: 10 max iterations. Rank4 is best when smallest

In the framework of a resubstitution error estimation method, Table VII shows the overall accuracy (sum of diagonal elements of the confusion matrix) between labeling results and reference ROIs. In this experiment, the performance of nontraditional algorithms (namely, MPAC, SEM, CSEM, and MSEM) is more competitive with those of traditional labeling approaches (namely, NP, ML and ICM-MAP-MRF) than in test case 1 (refer to Table III).

To compare the generalization capabilities of competing classifiers, Table VIII shows the maximum sum (after reshuffling) of diagonal elements of the overlapping area matrix computed between the reference cluster map x_i^* (generated by the ELBG vector quantizer) with the corresponding submap $x_i \subseteq x$, $i = 1, 2$ with $L = 12$ (see Section VI-C). In Table VIII, where ML shows the worst performance (as expected), the labeling fidelities to multiple cluster maps of output results provided by purely semisupervised algorithms SEM1, CSEM1, and MSEM1, as well as their partially semisupervised implementations SEM2, CSEM2, and MSEM2, are superior to those of the other labeling



(a) (b)

Fig. 4. (a) Test case 2. MPAC classification of the seven-band Landsat TM image, number of classes $L = 12$, shown in pseudocolors (refer to footnote 7). (b) Test case 2. MSEM1 classification of the seven-band Landsat TM image, number of classes $L = 12$, shown in pseudocolors (refer to footnote 7).

approaches, which is consistent in part with test case 1 (refer to Table IV). It is noteworthy that in line with theoretical expectations, MPAC (which is prone to detecting artifacts) performs more poorly in the less fragmented test case 2 than in test case 1. In test case 2, MPAC performs even worse than plug-in NP.

Overall, in line with theoretical expectations (refer to Section VI-B) and in line with test case 1 (refer to Section VII-A), the slight correlation between Rank5 (from resubstitution) and Rank6 (from generalization) reveals the presence of the Hughes phenomenon.

To investigate spatial fidelity of segmentation results to reference data, Table IX reports the mean of the edge map difference computed between a 4-adjaceny edge map extracted from the system's output map and the one extracted from every reference cluster map. In contrast with results shown in Table VIII, Table IX reveals that although SEM1, CSEM1 and MSEM1 perform better than ML (in line with theoretical expectations), they are ranked average in preserving genuine but small image details irrespective of their labeling. These clustering algorithms are outperformed by MPAC, which also performs better than NP (in line with theoretical expectations). Single-scale MRF-based contextual algorithms, ICM-MAP-MRF and CSEM, perform better than in test case 1 (refer to Table V), in line with theoretical expectations, which proves the strong application-dependency of MRF-based image mapping approaches on the optimization of class-specific clique potentials. The Spearman correlation value between Rank6 and Rank7 is 0.257, revealing poor agreement [19] (which justifies the separate, independent computation of indexes of labeling and segmentation fidelity of a map to reference data pursued by DAMA).

Computation time of the labeling algorithms is reported in Table X. These results are in line with those shown in Table VI.

Overall, conclusions in test case 2 seem rather consistent with those of test case 1 and with theoretical expectations about the algorithms' potential utilities.

VIII. Result Assessment

In the (subjective) assessment of quantitative experimental results proposed in this section, the evaluation criterion proposed in [12], where Zamperoni considers any new image processing algorithm worth disseminating among a broad audience if it may enrich a commercial image processing software toolbox, is taken into consideration.

TABLE VII
Test Case 2. Overall Resubstitution Accuracy (Sum of Diagonal Elements of the Confusion Matrix) Between Labeling Results and Reference Data (ROIs). Number of Label Types (= number of reference ROIs) = 12

Classifier	Overall resubstitution accuracy %	Rank5
NP	99.3	2.5
ML	99.8	1
ICM-MAP-MRF	99.3	2.5
MPAC	96.1	10
SEM1*	96.7	6.5
SEM2**	96.8	5
CSEM1*	96.6	8
CSEM2**	96.7	6.5
MSEM1*	97.3	4
MSEM2**	96.4	9

$*$: without supervised (training) samples. $**$: with supervised (training) samples. Rank5 is best when smallest

Let us collect results of test cases 1 and 2 in Table XI, where column Total (*Tot.*, best when smallest) is computed as $Rank1 + \ldots + Rank8$. *Score1* is the rank of column *Total*. Column *Accuracy* (*Acc.*, best when smallest) is computed as $Rank1 + Rank2 + Rank3 + Rank5 + Rank6 + Rank7$, i.e., *Accuracy* ignores the computational costs of the compared algorithms and accounts for both learning and generalization ability. *Score2* is the rank of column *Accuracy*. Column *Generalization ability* (*Gen.Ab.*, best when smallest) is computed as $Rank2 + Rank3 + Rank6 + Rank7$. *Score3* is the rank of column *Gen.Ab*. It is noteworthy that *Score3* is closely related to map quality indexes $Rank2$ and $Rank6$, which seem highly

TABLE VIII
TEST CASE 2. RESULTS OBTAINED BY THE DAMA STRATEGY FOR MAP QUALITY ASSESSMENT

Classifier	Standard value of the overlapping area (%) on reference cluster map 1 (sample mean = 49.2950, sample std = 6.5343)	Standard value of the overlapping area (%) on reference cluster map 2 (sample mean = 59.7540, sample std = 3.9556)	Average of standard mean values 1 and 2	Rank6
NP	-0.5043 (46.00)	0.7751 (62.82)	0.1354	8
ML	-1.9995 (36.23)	-2.5215 (49.78)	-2.2605	10
ICM-MAP-MRF	-0.8379 (43.82)	1.1442 (64.28)	0.1532	7
MPAC	-0.8211 (43.93)	-0.4131 (58.12)	-0.6171	9
SEM1*	0.7063 (53.91)	0.2998 (60.94)	0.5031	3
SEM2**	0.6910 (53.81)	0.3378 (61.09)	0.5144	2
CSEM1*	0.3681 (51.70)	0.1128 (60.20)	0.2404	5
CSEM2**	0.3788 (51.77)	0.0824 (60.08)	0.2306	6
MSEM1*	0.8103 (54.59)	0.4920 (61.70)	0.6512	1
MSEM2**	1.2082 (57.19)	-0.3094 (58.53)	0.4494	4

$*$: without supervised (training) samples. $**$: with supervised (training) samples. Rank6 is best when smallest

TABLE IX
TEST CASE 2. MEAN AND STANDARD DEVIATION OF THE DIFFERENCE EDGE MAP COMPUTED BETWEEN THE TWO EDGE MAPS MADE FROM x_i^* AND x_i, $i = 1,2$

Classifier	Standard value of the mean (in [0,4]) of edge map difference 1 (sample mean = 0.6065, sample std = 0.0266)	Standard value of the mean (in [0,4]) of edge map difference 2 (sample mean = 0.5575, sample std = 0.0269)	Average of standard mean values 1 and 2	Rank7
NP	-0.2065 (0.601)	-1.2440 (0.524)	-0.7253	3
ML	-0.3943 (0.596)	1.3554 (0.594)	0.4806	9
ICM-MAP-MRF	-0.8074 (0.585)	-1.6896 (0.512)	-1.2485	2
MPAC	-2.1592 (0.549)	-0.9481 (0.531)	-1.5716	1
SEM1*	-0.2816 (0.614)	0.3899 (0.568)	0.3358	6
SEM2**	-0.2065 (0.612)	0.2042 (0.563)	0.2054	4
CSEM1*	0.4694 (0.619)	0.4642 (0.570)	0.4668	8
CSEM2**	0.4694 (0.619)	0.4271 (0.569)	0.4482	7
MSEM1*	0.5069 (0.620)	-0.0186 (0.557)	0.2442	5
MSEM2**	1.6335 (0.650)	1.0955 (0.587)	1.3645	10

$*$: without supervised (training) samples. $**$: with supervised (training) samples. Rank7 is best when smallest

correlated to the qualitative map assessment criteria adopted by human photointerpreters featuring high generalization capability.

The arbitrary and problem-specific nature of map quality measures *Score1*, *Score2*, and *Score3* does not allow the reaching of any final conclusion about the accuracy and efficiency of the competing classifiers involved in the comparison. Nonetheless, the analysis of Table XI yields some relative (subjective) conclusions about the potential usability of the tested classifiers in dealing with the badly posed classification of piecewise constant or slowly varying color images (i.e., where texture information is negligible) whose fine image details are captured by multiple reference cluster maps. These relative conclusions can be considered interesting as they are based on weak (arbitrary, subjective), but numerous measures of image mapping quality that reasonably approximate the real-world

TABLE X
COMPUTATION TIMES OF THE INDUCTIVE LEARNING ALGORITHMS IN THE LANDSAT IMAGE TEST CASE

Classifier*	Computation time	Rank8
NP	23s	1
ML	1m 31s	2
ICM-MAP-MRF **	15m 50s	3
MPAC**	1h 40m 10s	8
SEM1, SEM2 **	32m 30s	4.5
CSEM1, CSEM2**	36m 30s	6.5
MSEM1, MSEM2**	4h 4m 30s	9.5

$*$: 12 label types, 20431 training pixels (2.6%). $**$: 10 max iterations. Rank8 is best when smallest

TABLE XI
SUMMARY OF EXPERIMENTAL RESULTS

	Rank1	Rank2	Rank3	Rank4	Rank5	Rank6	Rank7	Rank8	Tot.	Score1	Acc.	Score2	Gen.Ab.	Score3
NP	1	10	9	1	2,5	8	3	1	35,5	2	33,5	6	30	9
ML	3	8	4	2	1	10	9	2	39	4	35	7,5	31	10
ICM-MAP-MRF	2	9	10	3	2,5	7	2	3	38,5	3	32,5	5	28	7,5
MPAC	10	1	1	8	10	9	1	8	48	7	32	4	12	3
SEM1	9	4	2	4,5	6,5	3	6	4,5	39,5	5	30,5	3	15	4
SEM2	5	5	3	4,5	5	2	4	4,5	33	1	24	1,5	14	1,5
CSEM1	8	6	7	6,5	8	5	8	6,5	55	10	42	10	26	6
CSEM2	7	7	8	6,5	6,5	6	7	6,5	54,5	9	41,5	9	28	7,5
MSEM1	6	3	5	9,5	4	1	5	9,5	43	6	24	1,5	14	1,5
MSEM2	4	2	6	9,5	9	4	10	9,5	54	8	35	7,5	22	5

Tot. (best when smallest) $= Rank1 + \ldots + Rank8$. *Score1* is the rank of column *Tot. Accuracy* (best when smallest) $= Rank1 + Rank2 + Rank3 + Rank5 + Rank6 + Rank7$. *Score2* is the rank of column Accuracy (best when smallest). *Generalization Ability* ($Gen.Ab.$, best when smallest) $= Rank2 + Rank3 + Rank6 + Rank7$. *Score3* is the rank of column *Gen.Ab.*

characteristics of new generation image mapping applications featuring little ground truth knowledge.

- Subjective but numerous measures of image mapping quality, collected as *Score2* and *Score3*, reveal that when computational costs are ignored (which may be reasonable in a technological scenario where processing speed increases dramatically each year):
  - Nontraditional semisupervised and/or multiscale labeling approaches (namely, SEM, CSEM, MSEM and MPAC) seem capable of guaranteeing image mapping performance superior (on average) to those of first-generation classifiers. MSEM is (on average) superior to or competitive with the other competing approaches, especially when considering map quality indexes *Rank2* and *Rank6*, which seem to be highly correlated to the qualitative map assessment criteria adopted by human photointerpreters (related to generalization capability). In particular, MSEM appears to be largely superior to CSEM, but only slightly superior to SEM, which is in clear contrast with results reported in [22] (where CSEM outperforms SEM in several image mapping tasks). It is noteworthy that theoretical limitations of MPAC (artefact generation, see Section IV-A), known from existing literature, are confirmed by these experimental results (e.g., MPAC labeling accuracy is low when dealing with the raw image featuring coarser spatial details, see Rank6).
  - Among the tested traditional algorithms, none is ranked high.
- Subjective but numerous measures of image mapping quality, collected as *Score1*, reveal that:
  - Among nontraditional labeling strategies, the noncontextual iterative semisupervised SEM classifier provides an interesting compromise between labeling and spatial fidelity of results to reference data, with ease of use and low computational overhead. Moreover, SEM features a rigorous statistical foundation (unlike MPAC, CSEM, and MSEM), it can be employed in either partially semisupervised or purely semisupervised (where labeled samples with full weight are not employed) learning modes, and it does not apply exclusively to (2-D) images.
  - Traditional plug-in classifiers, namely, ML and NP, provide an acceptable tradeoff between labeling and spatial fidelity of results to reference data, ease of use and computational costs. This consideration justifies their diffusion in commercial image processing software toolboxes [42].

To summarize, results collected by DAMA (refer to *Score1*, *Score2*, and *Score3* in Table XI) in two badly posed image classification problems:

1) show that, although they rely on heavy class-conditional normal distribution assumptions that are not true in many real-world problems (causing a deterioration in the error rate for labeled samples), semisupervised classifiers using EM can be very powerful in practice;
2) appear always consistent with both theoretical considerations and subjective (perceptual) evaluations of output maps by expert photointerpreters (whereas results reported by optimistically biased resubstitution errors, see $Rank1$ and $Rank5$, appear to be counter-intuitive). On an *a posteriori* basis, the consistency between DAMA's map quality measures, qualitative evaluations by photointerpreters and several theoretical considerations appears to confirm the efficacy of the adopted DAMA strategy.

IX. CONCLUSIONS

The original heuristic multiscale semisupervised MSEM algorithm is proposed as a potential improvement over existing nonstandard data mapping systems, namely, SEM, CSEM, and MPAC, in the badly posed classification of piecewise constant or slowly varying color images (where texture information is negligible). To adequately compare MSEM against competing standard and nonstandard data mapping systems, subjective but numerous quantitative map quality measures are collected in badly posed image classification experiments where small image details are captured by multiple reference cluster maps. Experimental results reveal that, overall, MSEM is competitive with or superior to the other competing mapping systems (refer to *Score2* and *Score3* in Table XI), with special regard to quality indexes *Rank2* and *Rank6* that are highly correlated to empirical map quality criteria adopted by expert photointerpreters, at the cost of a computational overhead three to six times superior to that of its most competitive alternative, SEM, which

outperforms CSEM (in contrast with [22]). Overall, these experiments confirm that although they rely on heavy class-conditional normal distribution assumptions that are not true in many real-world problems, semisupervised classifiers using EM can be very powerful in practice.

Additional realistic, useful and relative conclusions about competing sample-based (i.e., context-insensitive) classification systems (unlike MSEM, MPAC, etc., refer to Table II), capable of processing generic 1-spatial dimensional sequences of multivariate data samples, can be derived from the collected quantitative results in the light of Zamperoni's recommendations [12]. In particular, SEM, which is provided with a rigorous statistical foundation and whose limitations are well known from existing literature [25], appears to be worthy of being disseminated among commercial data processing all-purpose software toolboxes, in that it is presumably useful to a broad audience dealing with pattern recognition problems, which may or may not involve images, either partially or purely semisupervised, either well- or badly posed. Finally, traditional noncontextual classifiers (namely, NP and ML), appear able to justify their diffusion among commercial data processing software toolboxes, owing to their theoretical simplicity, acceptable performance and competitive computational load.

As a future development of this work, a semisupervised learning approach, inspired by [2], should be combined with the supervised EM-based learning strategy developed for RBF network classifiers, proposed in [62].

REFERENCES

[1] A. Mojsilovic, J. Kovacevic, J. Hu, R. Safranek, and S. K. Ganapathy, "Matching and retrieval based on the vocabulary and grammar of color patterns," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 38–54, Sep. 2000.

[2] Q. Jackson and D. Landgrebe, "An adaptive classifier design for high-dimensional data analysis with a limited training data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 12, pp. 2664–2679, Dec. 2001.

[3] C. Kervrann and F. Heitz, "A Markov random field model-based approach to unsupervised texture segmentation using local and global spatial statistics," *IEEE Trans. Image Process.*, vol. 4, no. 6, pp. 856–862, Jun. 1995.

[4] T. N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Trans. Signal Process.*, vol. 3, no. 2, pp. 162–177, Feb. 1992.

[5] L. Bruzzone and D. Fernández Prieto, "A technique for the selection of kernel-function parameters in RBF neural networks for classification of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 2, pp. 1179–1184, Mar. 1999.

[6] M. Sgrenzaroli, A. Baraldi, H. Eva, G. De Grandi, and F. Achard, "Contextual clustering for image labeling: an application to degraded forest assessment in Landsat TM images of the Brazilian Amazon," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 8, pp. 1833–1847, Aug. 2002.

[7] G. Patanè and M. Russo, "The enhanced-LBG algorithm," *Neural Netw.*, vol. 14, no. 9, pp. 1219–1237, 2001.

[8] C. Liu and H. Wechsler, "A shape- and texture-based enhanced Fisher classifier for face recognition," *IEEE Trans. Image Process.*, vol. 10, no. 4, pp. 598–608, Apr. 2001.

[9] F. J. Cortijo and N. Perez de la Blanca, "Improving classical contextual classifications," *Int. J. Remote Sens.*, vol. 19, no. 8, pp. 1591–1613, 1998.

[10] J. T. Morgan, A. Henneguelle, J. Ham, J. Ghosh, and M. M. Crawford, "Adaptive feature spaces for land cover classification with limited ground truth," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 18, no. 5, pp. 777–799, 2004.

[11] E. Binaghi, I. Gallo, and M. Pepe, "A cognitive pyramid for contextual classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 12, pp. 2906–2922, Dec. 2003.

[12] P. Zamperoni, "Plus ça va, moins ça va," *Pattern Recognit. Lett.*, vol. 17, no. 7, pp. 671–677, 1996.

[13] R. C. Jain and T. O. Binford, "Ignorance, myopia and naivete' in computer vision systems," *Comput. Vis., Graph., Image Process.: Image Understanding*, vol. 53, pp. 112–117, 1991.

[14] M. Kunt, "Comments on 'Dialogue,' a series of articles generated by the paper entitled 'Ignorance, myopia and naivete' in computer vision systems'," *Comput. Vis., Graphics, Image Process.: Image Understanding*, vol. 54, pp. 428–429, 1991.

[15] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[16] V. Cherkassky and F. Mulier, *Learning From Data: Concepts, Theory, and Methods*. New York: Wiley, 1998.

[17] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[18] R. Kothari and V. Jain, "Learning from labeled and unlabeled data using a minimal number of queries," *IEEE Trans. Neural Netw.*, vol. 14, no. 6, pp. 1496–1505, Nov. 2003.

[19] R. G. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data*. Boca Raton, FL: Lewis, 1999.

[20] M. P. Buchheim and T. M. Lillesand, "Semi-automated training field extraction and analysis for efficient digital image classification," *Photogramm. Eng. Remote Sens.*, vol. 55, no. 9, pp. 1347–1355, 1989.

[21] Q. Jackson and D. Landgrebe, "An adaptive method for combined covariance estimation and classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 5, pp. 1082–1087, May 2002.

[22] Q. Jackson and D. A. Landgrebe, "Adaptive Bayesian contextual classification based on a Markov random field," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2454–2463, Nov. 2002.

[23] E. Backer and A. K. Jain, "A clustering performance measure based on fuzzy set decomposition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-3, no. 1, pp. 66–75, Jan. 1981.

[24] T. Lillesand and R. Kiefer, *Remote Sensing and Image Interpretation*, 3rd ed. New York: Wiley, 1994.

[25] M. M. Durat and D. A. Landgrebe, "A cost-effective semisupervised classifier approach with kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 1, pp. 264–270, Jan. 2004.

[26] L. Bruzzone, F. Roli, and S. Serpico, "Structured neural networks for signal classification," *Signal Process.*, vol. 64, pp. 271–290, 1998.

[27] A. Baraldi, P. Blonda, F. Parmiggiani, and G. Satalino, "Contextual clustering for image segmentation," *Opt. Eng.*, vol. 39, no. 4, pp. 1–17, Apr. 2000.

[28] A. Baraldi, M. Sgrenzaroli, and P. Smits, "Contextual clustering with label backtracking in remotely sensed image applications," in *Geospatial Pattern Recognit.*, E. Binaghi, P. Brivio, and S. Serpico, Eds. Kerala, India: Research Signpost/Transworld Research, Apr. 2002, in press.

[29] T. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[30] M. J. Barnsley and S. L. Barr, "Inferring urban land use from satellite sensor images using kernel-based spatial reclassification," *Photogramm. Eng. Remote Sens.*, vol. 62, no. 8, pp. 949–958, Aug. 1996.

[31] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, Feb. 2001.

[32] I. Kanellopoulos and G. G. Wilkinson, "Strategies and best practice for neural network image classification," *Int. J. Remote Sens.*, vol. 18, no. 4, pp. 711–725, 1997.

[33] E. J. Kaminsky, H. Barad, and W. Brown, "Textural neural network and version space classifers for remote sensing," *Int. J. Remote Sens.*, vol. 18, no. 4, pp. 741–762, 1997.

[34] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," *Pattern Recognit.*, vol. 24, no. 12, pp. 1167–1186, 1991.

[35] M. Ceccarelli and A. Petrosino, "Feature adaptive classifiers for SAR image segmentation," *Neurocomputing*, vol. 14, pp. 345–363, 1997.

[36] N. Petkov, "Biologically motivated image classification system," in *Real-Time Imaging*, P. A. Laplante and A. D. Stoyenko, Eds. New York: IEEE Press, 1996, pp. 195–223.

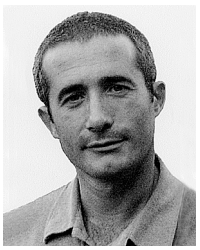
[37] L. Bruzzone and D. Fernández Prieto, "A technique for the selection of kernel-function parameters in RBF neural networks for classification of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 2, pp. 1179–1184, Feb. 1999.

[38] L. Bruzzone, D. Fernández Prieto, and S. B. Serpico, "A neural-statistical approach to multitemporal and multisource remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1350–1359, Mar. 1999.

[39] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. COM-31, no. 4, pp. 532–540, Apr. 1983.

[40] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge, 1997.

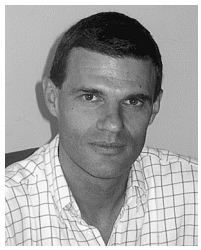
[41] J. S. De Bonet, "Novel Statistical Multiresolution Techniques for Imgage Synthesis, Discrimination, and Recognition," Ph.D. dissertation, MIT, Cambridge, MA, 1997.
[42] *"ENVI User's Guide,"* Research Systems, Inc., 2003.
[43] R. Haralick and L. Shapiro, "Survey of image segmentation techniques," *Comput. Vis., Graph., Image Process.*, vol. 29, pp. 100–132, 1985.
[44] Y. Jhung and P. H. Swain, "Bayesian contextual classification based on modified M-estimates and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 1, pp. 67–75, Jan. 1996.
[45] S. Krishnamachari and R. Chellappa, "Multiresolution Gauss-Markov random filed models for texture segmentation," *IEEE Trans. Image Process.*, vol. 6, no. 2, pp. 251–267, Feb. 1997.
[46] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Jun. 1984.
[47] J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Statist. Soc. B*, vol. 48, no. 3, pp. 259–302, 1986.
[48] M. Walessa and M. Datcu, "Model-based despeckling and information extraction from SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 5, pp. 2258–2269, Sep. 2000.
[49] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 2, pp. 98–110, Apr. 1993.
[50] A. Baraldi and P. Blonda, "A survey on fuzzy neural networks for pattern recognition: part I," *IEEE Trans. Syst., Man , Cybern. B, Cybern.*, vol. 29, no. 6, pp. 778–785, Dec. 1999.
[51] ——, "A survey on fuzzy neural networks for pattern recognition: Part II," *IIEEE Trans. Syst., Man , Cybern. B, Cybern.*, vol. 29, no. 6, pp. 786–801, Dec. 1999.
[52] R. N. Davè and R. Krishnapuram, "Robust clustering method: a unified view," *IEEE Trans. Fuzzy Syst.*, vol. 5, no. 2, pp. 270–293, 1997.
[53] A. H. Schistad Solberg, T. Taxt, and A. K. Jain, "A Markov random field model for classification of multisource satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 1, pp. 100–113, Jan. 1996.
[54] L. Delves, R. Wilkinson, C. Oliver, and R. White, "Comparing the performance of SAR image segmentation algorithms," *Int. J. Remote Sens.*, vol. 13, no. 11, pp. 2121–2149, 1992.
[55] L. Prechelt, "A quantitative study of experimental evaluations of neural network learning algorithms: current research practice," *Neural Netw.*, vol. 9, no. 3, pp. 457–462, 1996.
[56] P. Lukowicz, E. Heinz, L. Prechelt, and W. Tichy, Experimental Evaluation in Computer Science: A Quantitative Study Univ. Karlsruhe, Karlsruhe, Germany, Tech. Rep. 17/94, 1994.
[57] A. K. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.
[58] M. R. Spiegel*, Statistics*. New York: McGraw Hill, 1961.
[59] A. Baraldi, L. Bruzzone, and P. Blonda, "Quality assessment of classification and cluster maps without ground truth knowledge," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 857–873, Apr. 2005.
[60] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Int. Joint Conf. Artific. Intell.*, 1995 [Online]. Available: http://www.robotics.stanford.edu/users/ronnyk/ronnyk-bib.html
[61] [Online]. Available: http://www.dfc-grss.org
[62] P. H. Swain and S. M. Davis*, Remote Sensing: The Quantitative Approach*. New York: McGraw-Hill, 1978.
[63] D. Specht, "Probabilistic neural networks," *Neural Netw.*, vol. 3, pp. 109–118, 1990.
[64] E. Rignot and R. Chellappa, "Segmentation of polarimetric synthetic aperture radar data," *IEEE Trans. Image Process.*, vol. 1, no. 3, pp. 281–300, Mar. 1992.
[65] J. C. Bezdek, T. R. Reichherzer, G. S. Lim, and Y. Attikiouzel, "Multiple-prototype classifier design," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 28, no. 1, pp. 67–79, Feb. 1998.



Andrea Baraldi was born in Modena, Italy, in 1963. He graduated in electronic engineering from the University of Bologna, Bologna, Italy, in 1989. His M.S. thesis focused on the development of unsupervised clustering algorithms for optical satellite imagery.

From 1989 to 1990, he was a Research Associate at CIOC-CNR, an Institute of the National Research Council (CNR), Bologna, and served in the army at the Istituto Geografico Militare in Florence, Florence, Italy, working on satellite image classifiers and GIS. As a Consultant at ESA-ESRIN in Frascati, Italy, he worked on object-oriented applications for GIS from 1991 to 1993. From December 1997 to June 1999, he was with the International Computer Science Institute, Berkeley, CA, with a Postdoctoral Fellowship in Artificial Intelligence. From 2000 to 2002, as a Postdoctorate Researcher, he joined the European Commission Joint Research Centre, Ispra, Italy, in the development and validation of classification algorithms applied to wide area radar mosaics of forest ecosystems. Since his M.S. thesis, he has continued his collaboration with ISAC-CNR, Bologna, and ISSIA-CNR, Bari, Italy. He is currently with the IPSC-SES unit of the Joint Research Centre, involved with optical and radar image interpretation for terrestrial surveillance and change detection. His main interests center on image segmentation and classification, with special emphasis on texture analysis and neural network applications in computer vision.

Mr. Baraldi has been an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS since 2001.



Lorenzo Bruzzone (S'95–M'99–SM'03) received the "Laurea" (M.S.) degree in electronic engineering (summa cum laude) and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1993 and 1998, respectively.

He is currently Head of the Remote Sensing Laboratory in the Department of Information and Communication Technology, University of Trento, Trento, Italy. He was a Postdoctoral Researcher (1998 to 2000) and then Assistant Professor and Associate Professor (2001 to 2005) at the at the University of Genoa. Since March 2005, he has been a Full Professor of telecommunications at the University of Trento, where he currently teaches remote sensing, pattern recognition and electrical communications. His current research interests are in the area of remote-sensing image processing and recognition (analysis of multitemporal data, feature selection, classification, regression, data fusion, and neural networks). He conducts and supervises research on these topics within the frameworks of several national and international projects. Since 1999, he has been appointed Evaluator of Project Proposal for the European Commission. He is the author (or coauthor) of more than 140 scientific publications, including journals, book chapters, and conference proceedings.

Dr. Bruzzone ranked first in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (Seattle, WA, July 1998). He was a recipient of the Recognition of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGARS) Best Reviewers in 1999 and a Guest Editor of a Special Issue of the IEEE TGARS on the subject of the analysis of multitemporal remote sensing images (November 2003). He was the General Chair and Co-Chair, respectively, of the First and Second IEEE International Workshop on the Analysis of Multi-temporal Remote-Sensing Images. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He is an Associate Editor of the IEEE TGARS. He is a referee for many international journals and has served on the Scientific Committees of several international conferences. He is a member of the Scientific Committee of the *India-Italy Center for Advanced Research*. He is a member of the *International Association for Pattern Recognition* (IAPR) and of the *Italian Association for Remote Sensing* (AIT).

Palma Blonda (M'93) received the doctoral degree in physics from the University of Bari, Bari, Italy, in 1980.

In 1984, she joined the Institute For Signal and Image Processing (IESI), Italian National Research Council (CNR), Bari. Her research interests include digital image processing, fuzzy logic and neural networks, soft computing applied to the integration, and classification of multi-source remote sensed data. She was recently involved in the project "Landslide Early Warning Integrated System (LEWIS)," EVG1-CT-2001-00055, founded by the European Comunity in the framework of Fifth PQ. In the project, her research activity focuses on the application of multisource data integration and classification techniques for the extraction of EO-detectable superficial changes of some landslide related factors to be used in early warning mapping.