Information Sciences 217 (2012) 39–55

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Fusion of supervised and unsupervised learning for improved classification of hyperspectral images

Naif Alajlan [a], Yakoub Bazi [a], Farid Melgani [b], Ronald R. Yager [c,d,*]

[a] ALISR Laboratory, College of Computer and Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia
[b] Dept. of Information Engineering and Computer Science, Univ. of Trento, Via Sommarive, 14, I-38123 Trento, Italy
[c] Machine Intelligence Institute, Iona College, New Rochelle, NY 10801, United States
[d] King Saud University, Riyadh 11543, Saudi Arabia

ARTICLE INFO

Article history:
Received 20 August 2011
Received in revised form 2 June 2012
Accepted 21 June 2012
Available online 10 July 2012

Keywords:
Hyperspectral images
Support vector machine
Fuzzy c-means
Markov Fisher Selector
Voting rules
Markov Random Field

ABSTRACT

In this paper, we introduce a novel framework for improved classification of hyperspectral images based on the combination of supervised and unsupervised learning paradigms. In particular, we propose to fuse the capabilities of the *support vector machine* classifier and the *fuzzy C-means* clustering algorithm. While the former is used to generate a spectral-based classification map, the latter is adopted to provide an ensemble of clustering maps. To reduce the computation complexity, the most representative spectral channels identified by the Markov Fisher Selector algorithm are used during the clustering process. Then, these maps are successively labeled via a pairwise relabeling procedure with respect to the pixel-based classification map using voting rules. To generate the final classification result, we propose to aggregate the obtained set of spectro-spatial maps through different fusion methods based on voting rules and Markov Random Field theory. Experimental results obtained on two hyperspectral images acquired by the reflective optics system imaging spectrometer and the airborne visible/infrared imaging spectrometer, respectively; confirm the promising capabilities of the proposed framework.

1. Introduction

Airborne hyperspectral sensors have the capability of providing detailed measurements of the earth surface with very-high-spectral resolutions. This makes them powerful when dealing with applications requiring discrimination between subtle differences in ground covers (e.g., plant type differentiation, material quantification and target detection). However, the large dimensional data spaces (up to several hundreds) generated by these sensors can result in degraded classification accuracies. This is due to the curse of dimensionality (Hughes effect) characterizing this type of data [3,9,25,33].

Over the past years, various solutions dealing with the classification of hyperspectral images have been proposed in the literature of remote sensing. In the following, we present a review of the most recent published works. In detail, Tarabalka et al. [36–38] developed different mixed spatial and spectral classification schemes to improve the classification accuracy with respect to standard spectral classification methods. In [36] they combined a spectral classification map generated by the support vector machine (SVM) classifier from hyperspectral images characterized by 103 and 220 bands, and an ISODATA/EM-based segmentation map using the majority-voting concept within spatially connected regions. In [37], they used a minimum spanning forest from region markers. These markers were selected as the most reliable classified pixels from the SVM map. In [38], they applied the Markov Random Field (MRF) method as a post-processing scheme to an SVM-based

* Corresponding author at: King Saud University, Riyadh 11543, Saudi Arabia.
E-mail addresses: najlan@ksu.edu.sa (N. Alajlan), ybazi@ksu.edu.sa (Y. Bazi), melgani@disi.unitn.it (F. Melgani), yager@panix.com (R.R. Yager).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.06.031

classification map. In order to preserve edge information, an edge detector is applied to the hyperspectral image. Waske et al. [41] analyzed the robustness of multiple classifier systems based on SVM and random feature selection. They investigated how the number of selected features and the size of the multiple classifier system could influence the classification accuracy. Bourennane et al. [7] proposed a tensorial method termed as adaptive multidimensional Wiener filtering that performs simultaneously denoising and spectral dimensionality reduction. Kalluri et al. [19] proposed to enhance the classification accuracy by combining spectral reflectance information with the spectral derivative. To tackle the over-dimensionality introduced by the addition of these spectral derivatives, different solutions based on a single classifier or multiple classifier systems were adopted. Ma et al. [23] developed different approaches based on the combination of local manifold learning and the k-nearest neighbor (kNN) classifier. Fereidoun et al. [13] proposed to reduce the dimensionality using a generalized nonlinear discriminant analysis technique (GNDA), which is a kernel version of the Fisher linear discriminant analysis (FLDA). Then the transformed data were analyzed by the relevance vector machine (RVM) classifier. Mura et al. [29] presented a technique based on independent component analysis (ICA) and extended morphological attribute profiles (EAPs). During the classification stage, two approaches based on simple vector stacking and decision fusion were considered, respectively. Bakos and Gamba [2] introduced a multistage hierarchical data processing approach that allows to combine the advantages of different processing chains. This combination process was made using a hierarchical hybrid decision tree architecture. Zhong and Wang [50] developed a classification method based on conditioned random field (CRF) with sparse higher order potentials specially designed to incorporate complex characteristics of hyperspectral images. Jun and Ghosh [18] proposed a spatially adaptive classification method based on Gaussian process maximum likelihood (GP-ML) model. In this method, a Gaussian random process indexed by spatial coordinates was used to model each band of a given class. Then these models were used to characterize each land cover class at a given location by a multivariate Gaussian distribution with specific parameters adapted for that location. Li et al. [21] proposed to integrate the spectral and spatial information in a Bayesian framework. They used a multinomial logistic regression algorithm to learn the posterior probability distributions from the spectral information. Then, contextual information is included using a multilevel logistic Markov–Gibbs MRF prior. Li et al. [22] proposed two kinds of spatial–contextual SVMs for hyperspectral image classification. The first technique based on the concept of MRFs, uses the spatial information in the original space. By contrast, the second technique uses the spatial information in the feature space. Finally, Zhang et al. [49] introduced a patch alignment framework to linearly combine spectral, texture and shape features in an optimal way and obtain a unified low-dimensional representation of these multiple features for addressing the classification problem.

In this paper, we introduce a novel framework for improved classification of hyperspectral images by combining supervised and unsupervised learning. In particular, we propose to fuse the capabilities of the state-of-the-art SVM classifier and the fuzzy-c-means (FCM) clustering algorithm. The SVM classifier is used to generate a spectral-based classification map, whereas FCM is adopted to provide an ensemble of segmentation maps. To this end, we run different FCM's on the most relevant spectral channels selected by the Markov Fisher Selector (MFS) algorithm. Then, a pairwise relabeling of the spatial connected regions within these maps is made with respect to the SVM-based classification map using voting rules. The outcome of this operation is a set of spectral and spatial classification maps termed as SVM-FCM classification maps. To generate the final classification result, we propose to aggregate further these SVM-FCM maps through different fusion methods based on voting rules and MRF theory. While the formers are pixel-based aggregation methods, the latter exploits the spatial and inter-image contextual information during the fusion process. Experimental results obtained on two hyperspectral images acquired by ROSIS-03 and the AVIRIS sensors, respectively, confirm the interesting capabilities of the proposed classification framework.

The paper is organized as follows: A detailed description of the proposed classification framework is presented in Sections 2 and 3, respectively. The experimental results are reported in Section 4. Finally, conclusions and future directions are drawn in Section 5.

2. Framework overview

Let us consider a hyperspectral remote sensing image $\mathbf{X}$ of size $M \times N \times d$, where M, N, and d represent the number of rows, columns and spectral channels, respectively. Let us also assume a learning set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ from $\mathbf{X}$ composed of n training feature vectors $\mathbf{x}_i$ of dimension d and $y_i \in \{\omega_1, \ldots, \omega_T\}$ are class labels. Each class ω_i ($i = 1, \ldots, T$) has n_i training samples and T represent the number of classes. Given this training set $\mathcal{D}$, we aim at generating an improved classification map of the image $\mathbf{X}$ by computing the unknown class labels of the remaining (test) samples. Fig. 1 shows the general flow chart of the proposed classification system. In the following next subsections, we present the foundations of our framework.

2.1. Classification with SVM

SVMs are among the most popular supervised kernel-based classifiers available in the literature. Compared to standard classification methods they rely on the margin maximization principle that makes them less sensitive to overfitting problems [15,40]. We refer the reader to [28] for a good review related to the utilization of SVMs in the context of remote sensing.

For binary classification problems, to each vector $\mathbf{x}_i$ we associate a target $y_i \in \{-1, +1\}$. In non-linear SVMs, data are mapped with a kernel function in a higher dimensional feature space, i.e., $\Phi(\mathbf{x}) \in \Re^{d'}$ ($d' > d$). The membership decision rule

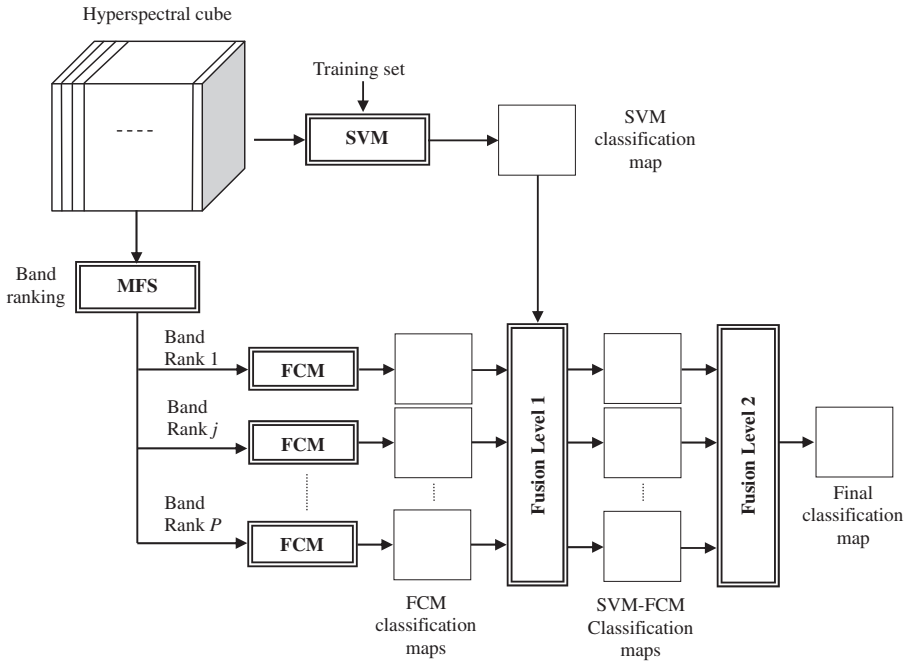

Fig. 1. General flow chart of the proposed framework.

is based on the function $sign[f(\mathbf{x})]$, where $f(\mathbf{x})$ represents the discriminant function associated with the hyperplane in the transformed space and is defined as:

$$f(\boldsymbol{x}) = \mathbf{w}^* \cdot \Phi(\mathbf{x}) + b^* \tag{1}$$

The optimal hyperplane defined by the weight vector $\mathbf{w}^* \in \Re^{d'}$ and the bias $b^* \in \Re$ is the one that minimizes a cost function that expresses a combination of two criteria: margin maximization and error minimization. It is expressed as:

$$\Psi(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{2}$$

This cost function minimization is subject to the following constraints:

$$y_i(\mathbf{w} \cdot \Phi(\mathbf{x}_i) + b) \geqslant 1 - \xi_i, \quad i = 1, 2, \ldots, n \tag{3}$$

and

$$\xi_i \geqslant 0, \quad i = 1, 2, \ldots, n \tag{4}$$

where the ξ_i's are the so-called slack variables introduced to account for non-separable data. The constant C represents a regularization parameter that controls the shape of the discriminant function, and consequently the decision boundary when data are non-separable. The above optimization problem can be reformulated through a Lagrange functional, for which the Lagrange multipliers can be found by means of a dual optimization leading to a Quadratic Programming (QP) solution, i.e.,

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{5}$$

under the constraints:

$$0 \leqslant \alpha_i \leqslant C \quad \text{for } i = 1, 2, \ldots, n \tag{6}$$

and

$$\sum_{i=1}^{n}\alpha_i y_i = 0 \tag{7}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_n]$ is a vector of Lagrange multipliers. The result is a discriminant function conveniently expressed as a function of the data in the original (lower) dimensional feature space $\boldsymbol{X}$:

$$f(\mathbf{x}) = \sum_{i \in S}\alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + b^* \tag{8}$$

where $K(\cdot,\cdot)$ is a kernel function. The set S is a subset of the indices $\{1,2,\ldots,n\}$ corresponding to the non-zero Lagrange multipliers α_i's, which define the so-called support vectors. The kernel $K(\cdot,\cdot)$ must satisfy the condition stated in Mercer's theorem so as to correspond to some type of inner product in the transformed (higher) dimensional feature space. A typical example of such kernels is represented by the following Gaussian function:

$$K(\mathbf{x}_i,\mathbf{x}) = \exp\left(-\gamma\|\mathbf{x}_i-\mathbf{x}\|^2\right) \tag{9}$$

where γ represents a parameter inversely proportional to the width of the Gaussian kernel.

The classification of hyperspectral remote sensing images often involves the simultaneous discrimination of numerous information classes. To this end, different multiclass classification strategies can be adopted [17,25]. In this paper, we apply the one-against-one scheme to generate the SVM-based classification map.

2.2. Band ranking with MFS

Compared with conventional multispectral data, hyperspectral data are characterized by a higher spectral resolution giving thus the opportunity to push further the information extraction capability. However, hyperspectral imagery involves a greater quantity of data to memorize and to process. In addition, they often exhibit redundant information, calling thus for opportune band (feature) selection algorithms [24].

Recently, Cheng et al. [10] proposed a fast feature selection algorithm suitable for multiclass classification problems. Its basic idea is to identify the best subset of features that maximizes the between-class distance while minimizing the within-class distance in a higher dimensional kernel space. By taking advantage of some special kernel functions such as the polynomial kernel, the general subset selection problem is then solved efficiently using MRF optimization techniques.

Let $\mathbf{S}_w$, $\mathbf{S}_b$ and $\mathbf{S}_t$ be the within-class, between-class and scatter matrices, respectively:

$$\mathbf{S}_w = \frac{1}{n}\sum_{j=1}^{T}\sum_{i=1}^{n_i}\left(\mathbf{x}_i^{(j)}-\boldsymbol{\mu}_j\right)\left(\mathbf{x}_i^{(j)}-\boldsymbol{\mu}_j\right)^T \tag{10}$$

$$\mathbf{S}_b = \frac{1}{n}\sum_{j=1}^{T}\left(\boldsymbol{\mu}_j-\boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_j-\boldsymbol{\mu}\right)^T \tag{11}$$

$$\mathbf{S}_t = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T \tag{12}$$

where $\mathbf{x}_i^{(j)}$ denotes the ith training sample in class ω_j, and $\boldsymbol{\mu}_j$ and $\boldsymbol{\mu}$ are the mean vectors of class ω_j and the complete training set, respectively.

After mapping data into the kernel space, the scatter matrices can be calculated in the kernel space. The trace of the within-, between-class, and total scatter matrices denoted by $\widetilde{\mathbf{S}}_w, \widetilde{S}_b$ and $\widetilde{\mathbf{S}}_t$ are then given as follows:

$$\mathrm{Tr}(\widetilde{\mathbf{S}}_w) = \frac{1}{n}\mathrm{Tr}(\mathbf{K}) - \sum_{i=1}^{T}\frac{1}{n_i}\mathrm{Sum}(\mathbf{K}^{(i)}) \tag{13}$$

$$\mathrm{Tr}(\widetilde{\mathbf{S}}_b) = \frac{1}{n}\sum_{i=1}^{T}\frac{1}{n_i}\mathrm{Sum}(\mathbf{K}^{(i)}) - \frac{1}{n^2}\mathrm{Sum}(\mathbf{K}) \tag{14}$$

$$Tr(\widetilde{\mathbf{S}}_t) = \frac{1}{n}\mathrm{Tr}(\mathbf{K}) - \frac{1}{n^2}\mathrm{Sum}(\mathbf{K}) \tag{15}$$

The operators $\mathrm{Sum}(\cdot)$ and $\mathrm{Tr}(\cdot)$ calculate, respectively, the summation of all elements and the trace of a matrix, and $\mathbf{K}$ and $\mathbf{K}^{(i)}$ are of size $n\times n$ and $n_i\times n_i$ matrices defined by:

$$\{\mathbf{K}\}_{k,l} = k(\mathbf{x}_k,\mathbf{x}_l), \quad \left\{\mathbf{K}^{(i)}\right\}_{u,v} = k\left(\mathbf{x}_u^{(i)},\mathbf{x}_v^{(i)}\right), \tag{16}$$

for $k, l\in\{1,\ldots,n\}$, $u,\ v\in\{1,\ldots,n_i\}$, and $i = 1,\ldots,T$.

The feature selector is denoted by $\boldsymbol{\alpha} = [\alpha_1,\ldots,\alpha_d]^T\in\{0,1\}^d$ with $\alpha_k = 1$ indicating the kth feature is chosen or 0 not-chosen. The selected features from the vector $\mathbf{x}$ are given by:

$$\mathbf{x}(\boldsymbol{\alpha}) = \mathbf{x}\odot\boldsymbol{\alpha} \tag{17}$$

where the operator $\odot$ represents the Schur product, which is an entrywise product. With feature selection, $\mathbf{K}$ and $\mathbf{K}^{(i)}$ become functions of $\boldsymbol{\alpha}$:

$$\{\mathbf{K}(\boldsymbol{\alpha})\}_{k,l} = k(\mathbf{x}_k\odot\boldsymbol{\alpha},\mathbf{x}_l\odot\boldsymbol{\alpha}), \quad \{\mathbf{K}^{(i)}(\boldsymbol{\alpha})\}_{u,v} = k\left(\mathbf{x}_u^{(i)}\odot\boldsymbol{\alpha},\mathbf{x}_v^{(i)}\odot\boldsymbol{\alpha}\right) \tag{18}$$

The traces of the scatter matrices become also functions of $\boldsymbol{\alpha}$ and are denoted by $\mathrm{Tr}(\widetilde{\mathbf{S}}_w)(\boldsymbol{\alpha})), \mathrm{Tr}(\widetilde{\mathbf{S}}_b)(\boldsymbol{\alpha})$ and $\mathrm{Tr}(\widetilde{\mathbf{S}}_t)(\boldsymbol{\alpha})$. The determination of the optimal vector $\boldsymbol{\alpha}^*$ is made by optimizing the following objective function:

$$\underset{\alpha\in\{0,1\}^d}{\operatorname{argamx}} \left\{ \operatorname{Tr}(\widetilde{\mathbf{S}}_b)(\boldsymbol{\alpha}) - \nu \operatorname{Tr}(\widetilde{\mathbf{S}}_t)(\boldsymbol{\alpha}) \right\} \tag{19}$$

where ν is a free parameter. We refer the reader to [10] for a detailed solution of this optimization problem.

The MFS algorithm has the peculiarity of assigning to each feature a coefficient indicating its importance. The higher the coefficient the more important the corresponding feature. We propose to use this information to identify the most relevant P spectral bands. Then on each individual spectral band, we apply an FCM algorithm to generate a set of P different clustering maps.

2.3. Clustering with FCM

The FCM algorithm [6,16] is a well known clustering method that is based on the minimization of the objective function:

$$J_m(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m(\mathbf{x}_k - \mathbf{v}_i)^T(\mathbf{x}_k - \mathbf{v}_i) \tag{20}$$

under the constraints:

$$\sum_{i=1}^{c} u_{ik} = 1 \tag{21}$$

and

$$u_{ik} \in [0,1] \tag{22}$$

where c represents the number of clusters. The $c \times n$ matrix $U = (u_{ik})$ is called the fuzzy partition matrix, where u_{ik} is the cluster membership of kth sample in the ith cluster, and m (usually set to 2) represents the fuzzy exponent that controls the amount of fuzziness.

A solution of the above optimization problem can be obtained via an iterative process where the fuzzy memberships and the clustering centers are iteratively updated via the following equations:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{(\mathbf{x}_k-\mathbf{v}_i)^T(\mathbf{x}_k-\mathbf{v}_i)}{(\mathbf{x}_k-\mathbf{v}_j)^T(\mathbf{x}_k-\mathbf{v}_j)}\right)^{1/(m-1)}}, \quad i = 1,\ldots,c, \quad \text{and} \quad k = 1,\ldots,n \tag{23}$$

$$\mathbf{v}_i = \frac{\sum_{k=1}^{n}(u_{ik})^m\mathbf{x}_k}{\sum_{k=1}^{n}(u_{ik})^m}, \quad i = 1,\ldots,c \tag{24}$$

The FCM algorithm terminates when there is no further change in the cluster centers. At convergence, each data sample is assigned to the cluster to which it has maximum membership.

In this work, we propose to use FCM for generating an ensemble of P different clustering maps. We run FCM P-times on the top relevant spectral channels obtained by the MFS algorithm. To reduce the computation complexity, at each time of the clustering process we consider only one spectral band as input to FCM. As no training data is used during the clustering process, the class-labels contained in each generated image are arbitrary. Therefore, a pairwise relabeling of this ensemble of maps with respect to the SVM-based classification map should be made before aggregation. Nonetheless, SVM usually leads to accurate classification results with respect to FCM as it is supervised. A direct alignment of the FCM maps with respect to the SVM map would result in poor results. Instead, we propose to label successively the spatially connected regions (segments) constructed by grouping the connected pixels in these clustered maps. Then, these segments are independently classified by employing decision fusion operators on the pixels belonging to the corresponding pixels in the SVM-based classification map. A description of this fusion mechanism is presented Section 3.1.

3. Information fusion

Information fusion is usually known as the process of fusing data coming from various sources by using a proper aggregation technique [20,44]. Over the past decades many operators have been proposed to fuse data information such as the order weighted average operator (OWA) [45,46], generalized OWA operator [47], Bonferroni mean operator [42,48] and power geometric operator [43]. In the context of remote sensing, fusion has been adopted to solve many issues related to classification [5,12,39], target recognition [32], change detection [26], and estimation of biophysical parameters [8].

In this framework, we implement the fusion process at two levels. The first level of fusion aims at generating an ensemble of SVM-FCM classification maps from the FCM clustered maps via a pairwise relabeling procedure with respect to the SVM classification map. For such purpose, we propose to explore the majority voting (MV) and weighted majority voting (WMV) rules, which are particular cases of the OWA operator introduced by Yager and Kacprzyk [45] and Yager [46,47]. The second level of fusion aims at generating the final classification map by further aggregating this ensemble of SVM-FCM maps. To

reach this objective, we propose two different aggregation schemes relying on pixel and contextual information, respectively. While the former relies also on voting rules, the latter exploits the power of MRF theory [11].

3.1. Generation of the SVM-FCM classification maps

According to the MV rule, a segment is assigned to the class ω_i $(i = 1,\ldots,T)$ if the number of pixels belonging to this class in the corresponding SVM map is the most frequent. For the WMV rule, we consider the sum of the posterior probabilities obtained by applying the Platt procedure to the SVM outputs during the classification phase [31]. The following algorithm shows the generation of the set of SVM-FCM classification maps.

Algorithm 1. SVM-FCM Maps Generation

Inputs:
- $\mathbf{X}$ – hyperspectral image
- $\mathbf{Y}^{SVM}$ – SVM-classification map
- P – Desired number of segmentation maps
- c – Number of clusters

Outputs:

1. $\mathbf{Y}_i$, $i = 1,\ldots,P$ – SVM-FCM classification maps
   Apply the MFS algorithm to image $\mathbf{X}$ and select the most relevant P spectral channels.
2. *for* $i = 1{:}P$
   - 2.1 Generate a clustering map by applying an FCM to the ith spectral channel selected by MFS.
   - 2.2. Construct a segmentation map by identifying the spatially connected regions (segments) in the ith clustered map.
   - 2.3. Classify successively the segments of the ith segmentation map to the class ω_j $(j = 1,\ldots,T)$ by applying the MV (WMV) rules to the corresponding pixels in the $\mathbf{Y}^{SVM}$ map.
3. end

3.2. Pixel-based information fusion

Under the spatial pixel independence assumption, the decision for each pixel is made independently on its neighboring. For the WMV rule, we propose to further weight the contribution of each SVM-FCM map during the fusion process. The availability of the training samples makes the computation of these weights particularly easy. If $(OA_i^{tr}, i = 1\ldots,P)$ is the overall accuracy on the training set for the ith SVM-FCM map. Then the weights are computed as follows:

$$\beta_i = \frac{OA_i^{tr}}{\sum_{j=1}^{P} OA_j^{tr}}, \quad \text{for } i = 1,\ldots,P, \quad \text{with } \sum_{i=1}^{P}\beta_i = 1. \tag{25}$$

The generation of the final classification map $\mathbf{Y}^*$ from the $(\mathbf{Y}_i, i = 1,\ldots,P)$ SVM-FCM maps by using the MV (WMV) rules is summarized in the following algorithm:

Algorithm 2. MV/WMV

Inputs:
- $\mathbf{Y}_i$, $i = 1,\ldots,P$ – SVM-FCM classification maps

Outputs:
- $\mathbf{Y}^*$ – final classification map

1. Compute the weights β_i associated to each SVM-FCM map according to (25).
2. *for* $k = 1{:}M$, $l = 1{:}N$
   - 2.1 Identify the class ω_j $(j = 1,\ldots,T)$ for the pixel y^*_{kl} by applying the MV (WMV) rules to the corresponding pixels y^j_{kl} in the SVM-FCM maps.
3. end

3.3. Contextual-based information fusion

The analysis of image pixels under spatial independence assumption may lead to inconsistencies mainly due to the presence of noise. Therefore, making a decision for a pixel by taking into account its neighborhood often represents an effective way for increasing the accuracy of the result. To this end, MRFs have proved a powerful and successful mathematical

framework as shown by various works dealing with different remote sensing problems [4,26,27,30,34]. In this work, we propose to adopt the MRF approach to perform the contextual fusion of the SVM-FCM maps.

To generate the final classification map $\mathbf{Y}^*$, given the SVM-FCM maps ($\mathbf{Y}_i$, $i = 1,2,\ldots,P$), one can resort to the maximum a posteriori probability (MAP) decision criterion:

$$P(\mathbf{Y}^*|\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_p) = \max_{\mathbf{Y}}\{P(\mathbf{Y}|\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_p)\} \tag{26}$$

The adoption of the MRF approach simplifies the complexity of this maximization problem by passing from a global model to a model of the local image properties. The combination of the MAP method with the MRF modeling makes the classification task equivalent to the minimization of a total energy function U_T expressed in the following relationship:

$$P(\mathbf{Y}^*|\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_p) = \frac{1}{Z}\exp\left[-U_T(\mathbf{Y},\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_p)\right] \tag{27}$$

where Z is a normalizing constant.

The total energy function U_T can be rewritten in terms of local energy functions U_{kl} using the concept of neighborhood [11]:

$$U_T(\mathbf{Y},\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_p) = \sum_{k=1}^{M}\sum_{l=1}^{N}U_{kl} \tag{28}$$

with

$$U_{kl} = U\left(y_{kl}, Y^{\mathcal{N}}(k,l), Y_1^{\mathcal{N}}(k,l),\ldots,Y_p^{\mathcal{N}}(k,l)\right) \tag{29}$$

where $Y^s(k,l)$ and $Y_i^s(k,l)$ stand for the set of labels of the pixels of the image $\mathbf{Y}$ and the images $\mathbf{Y}_i$ ($i = 1,2,\ldots,P$), respectively, in a predefined neighborhood system $\mathcal{N}$ associated with pixel (k,l).

The minimization of (29) can be carried out by means of different algorithms such as simulated annealing (SA), the maximizer of posterior marginals (MPMs), and the iterated conditional modes (ICMs) algorithms [11]. In this work, the ICM algorithm is adopted since it represents a simple and computationally moderate solution to optimize the MRF-MAP estimates, for it converges to a local, but usually good, minimum of the energy function. The ICM consists of minimizing iteratively the total energy function U_T through a pixel-based scheme until convergence is reached (i.e., where the pixel labels do not change much).

We propose to decompose the local energy function U_{kl} into two kinds of sources of contextual information, which contribute to the optimization process. The first is the spatial contextual information source, which defines the spatial correlation in image $\mathbf{Y}$ between the label of pixel (k,l) and the labels of its neighbors. The second is the inter-image information sources, which express the relationship between the image $\mathbf{Y}$ and each of the classification maps $\mathbf{Y}_i$ ($i = 1,2,\ldots,P$). Accordingly, the local energy function U_{mn} to be minimized for the pixel (m, n) can be written as follows:

$$U_{kl} = \beta_{SP}\cdot U_{SP}\left(y_{kl}, Y^{\mathcal{N}}(k,l)\right) + \sum_{i=1}^{P}\beta_i\cdot U_{II}\left(y_{kl}, Y_i^{\mathcal{N}}(k,l)\right) \tag{30}$$

where U_{SP} and U_{II} refer to the spatial and inter-image energy functions, respectively, while β_{SP} and β_i ($i = 1,2,\ldots,P$) represent the spatial and inter-image parameters, respectively. The weights β_i aim at controlling the contribution of the SVM-FCM maps during the fusion process. According to the values of these weights, we define two MRF fusion schemes. If they are set to one, we obtain an MV-MRF fusion scheme. By contrast, if these weights are computed according to (25) we obtain an WMV-MRF fusion scheme.

The neighborhood system adopted to define the two kinds of energy functions required to compute the local energy function in (30) is based on a second-order neighborhood. Based on this neighborhood system, the spatial energy function can be expressed as:

$$U_{SP}\left(y_{kl}, Y^{\mathcal{N}}(k,l)\right) = -\sum_{y_{kl}\in\mathbf{Y}^{\mathcal{N}}(k,l)} I(y_{kl}, y_{pq}) \tag{31}$$

where I is the Indicator function which allows us to count the number of occurrences of y_{kl} in $\mathbf{Y}^{\mathcal{N}}$ (the spatial part of $\mathcal{N}$) and is defined as:

$$I(y_{kl}, y_{pq}) = \begin{cases} 1, & \text{if } y_{kl} = y_{pq} \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

The inter-image energy function characterizing the spatial correlation between the image $\mathbf{Y}$ and the images $\mathbf{Y}_i$ ($i = 1,2,\ldots,P$) is given by:

$$U_{II}\left(y_{kl}, Y_i^{\mathcal{N}}(k,l)\right) = -\sum_{Y_i(p,q)\in Y_i^{\mathcal{N}}(k,l)} I(y_{kl}, Y_i(p,q)) \tag{33}$$

In the following, we present the algorithm of the proposed MRF fusion method:

Algorithm 3. MV-MRF/WMV-MRF

Inputs:
$\mathbf{Y}_i$, $i = 1,\ldots,P$ – SVM-FCM classification maps
β_{SP} – Spatial regularization parameter
Iter – Number of iterations
Outputs:
$\mathbf{Y}^*$ – final classification map
1. Compute the weights β_i associated to each SVM-FCM map according to (25).
2. Initialize $\mathbf{Y}$ by minimizing for each pixel (k, l) the local energy function U_{kl} defined in (30) without the spatial energy term (i.e., by setting $\beta_{SP} = 0$).
3. *for* $j = 1 : Iter$
4. Update $\mathbf{Y}$ by minimizing for each pixel (k, l) the local energy function U_{kl} defined in (30) including the spatial energy term β_{SP}.
5. end

4. Experimental results

4.1. Dataset description and performance evaluation

The first hyperspectral real dataset represents an urban area that was acquired by the ROSIS-03 optical sensor in 2002 over the University of Pavia, Italy. The image is 610 × 340 pixels, with a spatial resolution of 1.3 m per pixel. The number of data channels in the original recorded image was 115 (with a spectral range from 0.43 to 0.86 ψm). The 12 noisiest channels have been removed, and the remaining 103 bands were used in experiments. A single spectral channel of this image is shown in Fig. 2a. The ground truth image (see Fig. 2b) contains nine classes of interest. Table 1a reports the detailed number of samples for each class.

The second hyperspectral real dataset represents a section of a scene acquired in 1992 over the Indian test Pines site in Northwestern Indiana by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor. The image has a spatial dimension of 145 × 145 pixels and a spatial resolution of 20 m per pixel. All available 220 spectral channels were used in the experiments. Fig. 3a shows a single spectral channel of the Indian Pine image. The ground-truth image used to assess the proposed approach consists of sixteen land-cover classes is shown in Fig. 3b. The detailed number of samples characterizing each class is reported in Table 1b.

To obtain reliable assessment of the classification results, the experiments were repeated five times with different training and test samples generated randomly from the ground truth image. The results on these five trials on the different test sets were thus averaged. For the Pavia dataset, 50 training samples were considered for each class as training and the remaining samples were used as test. Concerning the Indiana dataset, 50 training samples were also considered for each class, expect for the minority classes *Alfalfa*, *Grass/Pasture mowed* and *Oats* where only 15 training samples were used. This means that 1.05% and 6.70% were used as training for Pavia and Indiana datasets, respectively.

The classification performances were obtained on a personal computer with a 1.86 GHz processor using Matlab 9 and are expressed in terms of class-by-class, overall (OA), average (AA) accuracies, standard deviation (σ_{OA}, σ_{AA}) in addition to the computation time. Furthermore, the statistical significance of differences between the accuracies achieved by the different classifiers was assessed using the *McNemar*'s test, which is based on the standardized normal test statistic [1,14]:

$$Z_{ij} = \frac{f_{ij} - f_{ji}}{\sqrt{f_{ij} + f_{ji}}} \tag{34}$$

where Z_{ij} measures the pairwise statistical significance of the difference between the accuracies of the ith and jth classifiers. f_{ij} stands for the number of samples classified correctly and wrongly by the ith and jth classifiers, respectively. Accordingly, f_{ij} and f_{ji} are the counts of classified samples on which the considered ith and jth classifiers disagree. At the commonly used 5% level of significance, the difference of accuracies between the ith and jth classifiers is said statistically significant if $|Z_{ij}| > 1.96$.

4.2. Results

4.2.1. Classification with SVM

We applied the SVM classifier directly in the original high dimensional space. During the training phase, the SVM parameters C and γ were selected according to a k-fold cross-validation (CV) procedure [35] first by randomly splitting the training data into k mutually exclusive subsets (folds) of equal size, then by training k times an SVM classifier modeled with

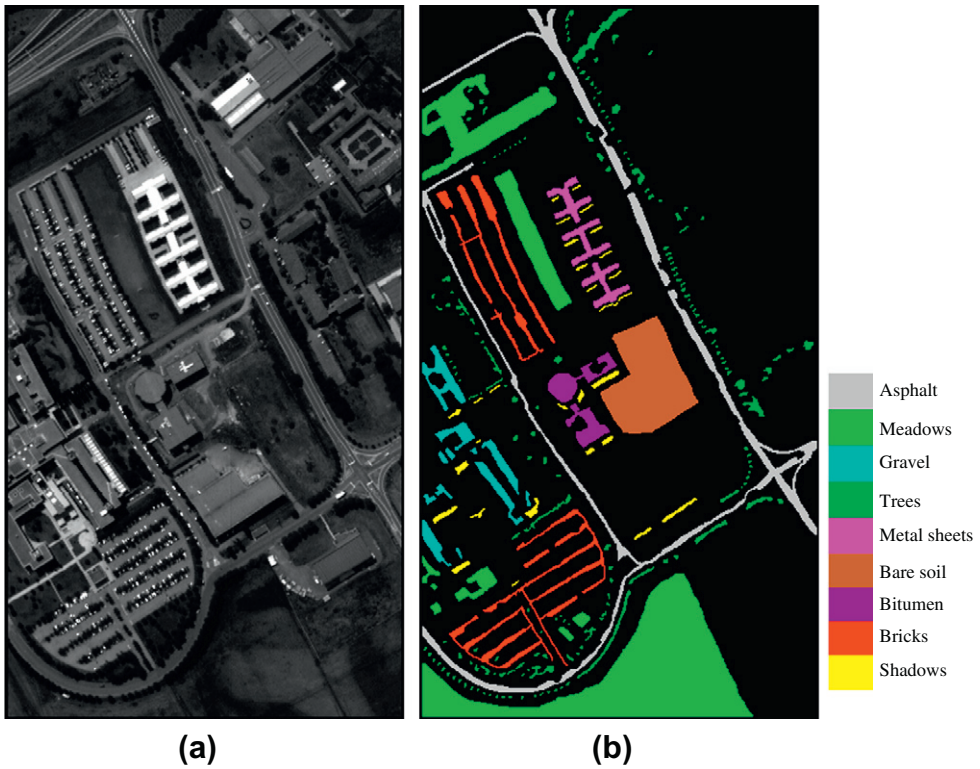

Fig. 2. Pavia dataset: (a) the 118th spectral band identified by the MFS algorithm as the most relevant one and (b) the ground truth image.

Table 1
Number of samples used in the experiments for (a) the Pavia and (b) the Indiana datasets.

Class	# Samples
(a)	
ω_1 – Asphalt	6631
ω_2 – Meadows	18,649
ω_3 – Gravel	2099
ω_4 – Trees	3064
ω_5 – Metal sheets	1345
ω_6 – Bare soil	5029
ω_7 – Bitumen	1330
ω_8 – Bricks	3682
ω_9 – Shadows	947
(b)	
ω_1 – Corn-no till	1434
ω_2 – Corn-min till	834
ω_3 – Corn	234
ω_4 – Soybeans-no till	968
ω_5 – Soybeans-min till	2468
ω_6 – Soybeans-clean till	614
ω_7 – Alfalfa	54
ω_8 – Grass/Pasture	497
ω_9 – Grass/Trees	747
ω_{10} – Grass/Pasture-mowed	26
ω_{11} – Hay-windrowed	489
ω_{12} – Oats	20
ω_{13} – Wheat	212
ω_{14} – Woods	1294
ω_{15} – Bldg–Grass–Tree–Drives	380
ω_{16} – Stone-steel towers	95

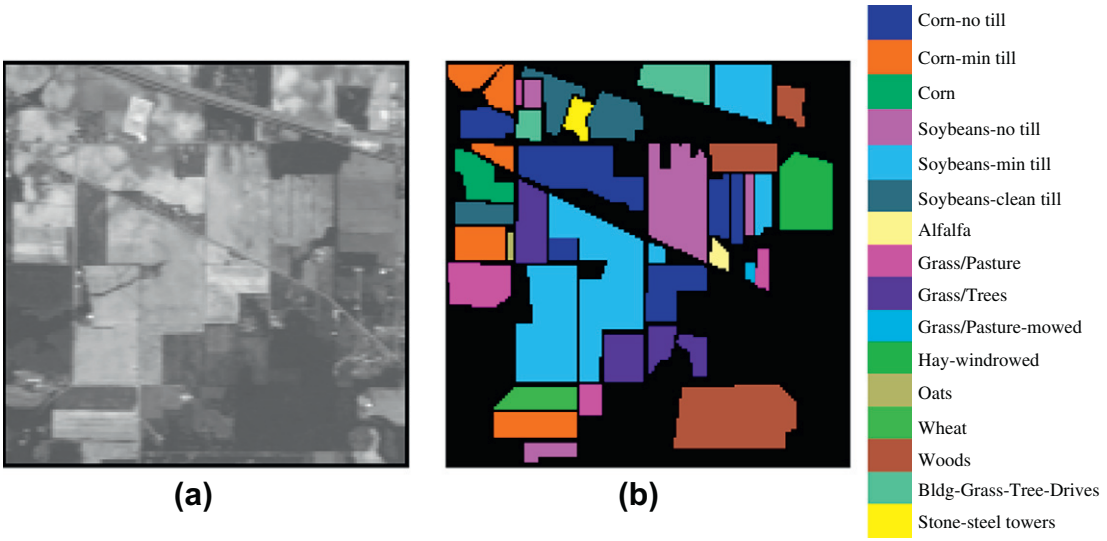

Fig. 3. Indiana dataset: (a) the 29th spectral band identified by the MFS algorithm as the most relevant one and (b) the ground truth image.

predefined values of C and γ. Each time we left out one of the subsets from training, and used the omitted subset only to obtain an estimate of the classification accuracy. From k times of training and accuracy computation, the average accuracy yielded a prediction of the classification accuracy of the considered SVM classifier. The best SVM classifier parameter values were chosen, to maximize such a prediction. In this experiment, we used a 3-fold CV by varying the values of C and γ in the range $[10^{-3}; 1000]$ and $[10^{-3}; 5]$, respectively. For Pavia, the corresponding OA and AA achieved on the test set were 84.34% and 89.00%, respectively. The worst class accuracy was obtained for the "ω_1 – *Asphalt*" class as it was equal to 76.97%. Concerning Indiana, the OA and AA were equal to 81.16% and 87.16%, respectively. The worst class accuracy was obtained for the "ω_5 – *Soybeans-min till*" class as it was equal to 68.86%. This experiment provided us with reference classification accuracies in order to quantify the capability of the proposed classification framework in improving these results.

4.2.2. Feature ranking with MFS

The application of the MFS algorithm for ranking the spectral channels yielded the distributions of the coefficients shown in Fig. 4. The relevant $P = 10$ spectral channels (used to construct the ensemble) for Pavia were {18, 17, 19, 20, 16, 21, 15, 22, 14, 23}, whereas they were {29, 28, 30, 27, 26, 32, 31, 25, 24, 33} for Indiana. By visual inspection, we observe that these bands are clean from noise. By contrast, the noisy bands were associated with lower coefficients. For example, for the Indiana dataset, the atmospheric and water absorption bands (104–108, 150–163–220–224) were assigned lower coefficients. It is interesting to notice that the MFS algorithm took only 0.13 and 0.18 s for ranking the features of the Pavia and the Indiana datasets, respectively.

4.2.3. Generation of the SVM-FCM maps

In order to generate the ensemble of clustering maps, we run the FCM algorithm P-times each time on one of the above selected spectral bands. Since the number of classes is known a priori for both datasets. A natural choice is to set the number of clusters to $c = 9$ and $c = 16$, respectively for Pavia and Indiana datasets, respectively. However, to increase further the diversity we allow the generation of over-clustered maps by randomly varying c in the range [9,14,16,21], respectively (i.e., we allowed up to five more clusters). After clustering, the spatially connected regions within these maps were assigned arbitrary labels via the function "*bwlabel*" of Matlab. Then, the generation of the of SVM-FCM maps is made by assigning these regions to the class labels by applying MV and WMV rules on the corresponding regions in the SVM-based classification map as shown in Fig. 5.

4.2.4. Classification with MV, WMV, MV-MRF and WMV-MRF

Tables 2 and 3 report the classification results obtained by the proposed fusion methods for Pavia and Indiana datasets, respectively. The Markovian fusion methods were applied by considering different values of the parameter β_{SP} in the range [1,10]. The obtained results did not change significantly for this range of values suggesting that the setting of this parameter within the considered range is not critical. However, it is interesting to notice that one could use the method proposed by Serpico and Moser [34] for an automatic setting of this parameter. The results reported in Tables 2 and 3 are for the value $\beta_{SP} = 1.5$. As can be seen, for Pavia dataset, the MV, WMV, MV-MRF and WMV-MRF methods yielded an (OA and AA) of (93.67% and 94.49%), (93.42% and 94.38%), (95.89% and 96.62%) and (95.99% and 96.75%), respectively. The corresponding

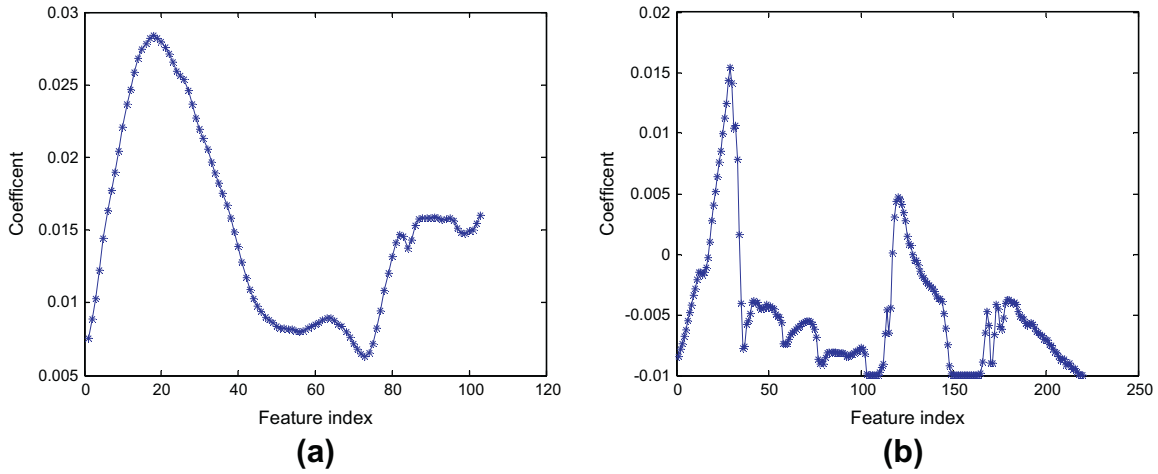

Fig. 4. Coefficients attributed by the MFS algorithm to the spectral channels for: (a) the Pavia and (b) the Indiana datasets.

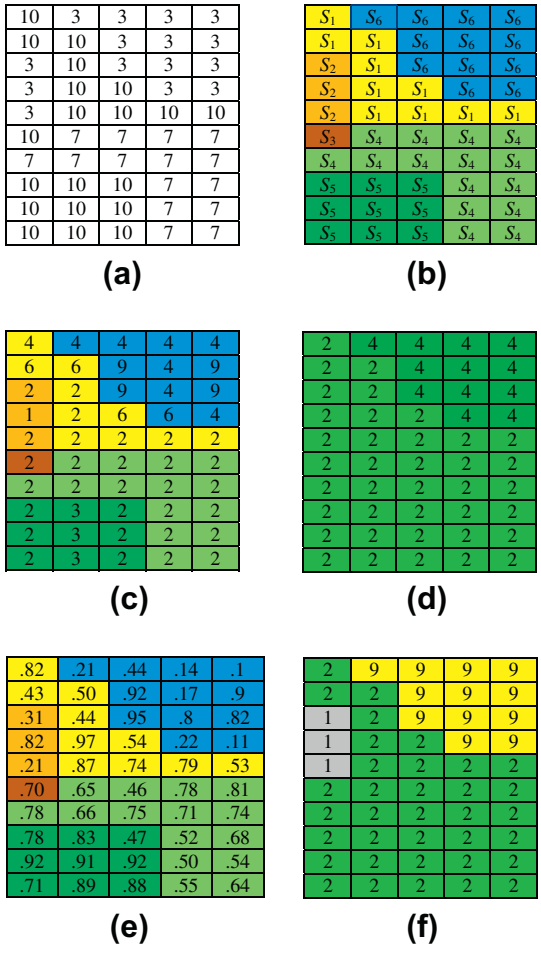

Fig. 5. Example of generation of an SVM-FCM map: (a) FCM result (the labels are arbitrary); (b) identified segments; (c) corresponding class labels generated by SVM; (e) result obtained by applying MV rule to the class labels obtained by SVM; (f) posterior probabilities of the class labels obtained by SVM; and (e) result obtained by applying the WMV rule to the posterior probabilities.

Table 2
Overall (OA), average (AA), standard deviations (σ) and class percentage accuracies achieved SVM, MV, WMV, MV-MRF and WMV-MRF classifiers on the test samples for the Pavia dataset. The total time (in seconds) required by each of them is also reported.

Class	SVM	MV	WMV	MV-MRF	WMV-MRF
ω_1	76.97	85.20	83.06	92.84	92.53
ω_2	82.17	95.51	95.33	95.97	**96.05**
ω_3	80.85	91.75	90.97	**94.62**	93.92
ω_4	95.93	92.05	92.26	92.63	93.78
ω_5	99.33	99.32	99.33	99.98	**100**
ω_6	88.09	95.39	96.22	98.06	**98.44**
ω_7	94.45	97.59	97.81	98.75	**99.15**
ω_8	83.23	95.08	95.92	98.38	**98.44**
ω_9	**99.95**	98.52	98.52	98.35	98.43
OA	84.34	93.67	93.42	95.89	**95.99**
σ_{OA}	1.35	1.57	1.76	1.72	1.88
AA	89.00	94.49	94.38	96.62	**96.75**
σ_{AA}	0.14	0.29	0.22	0.37	0.35
Time (s)	**17**	628	628	1360	1120

Table 3
Overall (OA), average (AA), standard deviations (σ) and class percentage accuracies achieved SVM, MV, WMV, MV-MRF and WMV-MRF classifiers on the test samples for the Indiana dataset. The total training time (in seconds) required by each of them is also reported.

Class	SVM	MV	WMV	MV-MRF	WMV-MRF
ω_1	77.93	84.69	83.35	87.65	86.83
ω_2	77.55	83.67	83.29	87.27	**87.83**
ω_3	93.36	94.67	95.65	**99.67**	99.56
ω_4	77.73	84.94	84.88	88.47	**88.58**
ω_5	68.86	77.81	80.10	84.44	**85.73**
ω_6	88.51	89.85	91.80	94.92	**96.70**
ω_7	74.87	82.56	83.07	89.23	**89.74**
ω_8	91.49	93.64	93.42	**96.24**	95.74
ω_9	94.66	97.81	97.76	98.96	**99.34**
ω_{10}	92.72	**98.18**	**98.18**	**98.18**	**98.18**
ω_{11}	98.54	98.86	99.04	99.63	**99.81**
ω_{12}	100	100	100	96.00	**100**
ω_{13}	99.62	99.62	99.62	99.75	**100**
ω_{14}	89.37	94.19	94.11	94.58	**94.59**
ω_{15}	76.54	84.24	86.78	93.51	**94.36**
ω_{16}	92.88	96.44	96.88	99.11	**100**
OA	81.16	86.93	87.48	90.67	**91.05**
σ_{OA}	0.82	0.97	0.95	1.21	1.13
AA	87.16	91.32	91.75	94.22	**94.81**
σ_{AA}	1.22	0.92	1.04	0.68	0.69
Time (s)	**27**	75	75	231	150

Table 4
Statistical significance of differences in classification accuracy between the six investigated classifiers expressed by means of the McNemar's test for (a) the Pavia and (b) the Indiana datasets. Statistically significant differences at the 5% level of significance ($|Z_{ij}| \geqslant 1.96$) are highlighted in bold face. Positive values indicate that the ith classifier has a classification accuracy higher than that of the jth one.

	MV	WMV	MV-MRF	WMV-MRF
(a)				
SVM	**−56.30**	**−54.87**	**−64.17**	**−64.71**
MV		**5.15**	**−26.29**	**−26.61**
WMV			**−27.99**	**−29.28**
MV-MRF				**−2.38**
(b)				
SVM	**−18.20**	**−19.84**	**−25.14**	**−26.07**
MV		**−4.08**	**−14.89**	**−15.65**
WMV			**−12.80**	**−14.33**
MV-MRF				**−2.74**

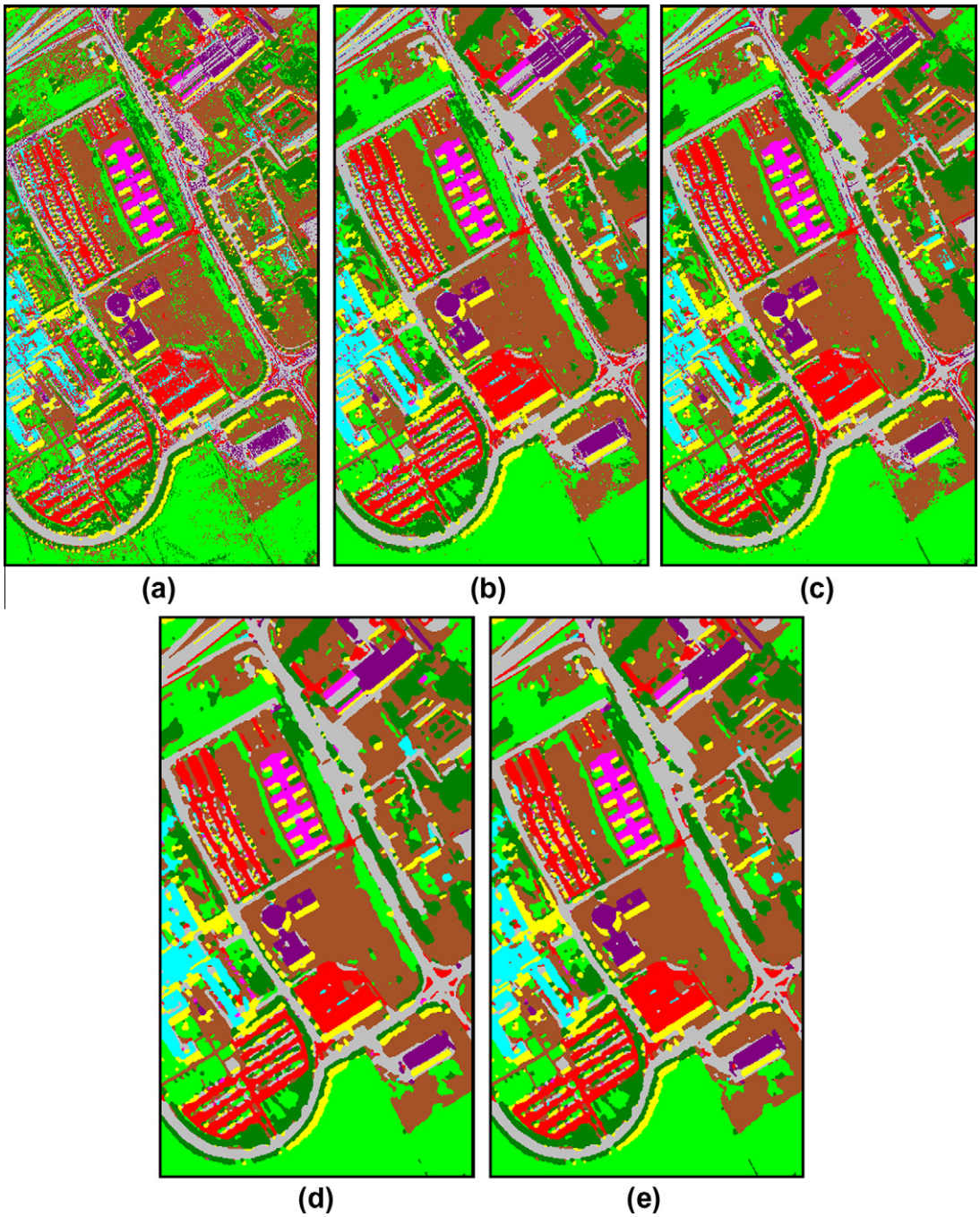

Fig. 6. Classification maps obtained by (a) SVM, (b) MV, (c) WMV, (d) MV-MRF and (e) WMV-MRF for the Pavia dataset.

gain in accuracy with respect to SVM was significant as it was equal to (9.33% and 5.49%), (9.08% and 5.38%), (11.55% and 7.62%) and (11.65% and 7.75%), respectively. Similar to SVM, the worst class accuracy for these fusion methods was obtained for "ω_1 – *Asphalt*" class and was equal to 85.20%, 83.06%, 92.84% and 92.53%, respectively. The corresponding gain in accuracy with respect to SVM was equal to 8.23%, 6.09%, 15.87% and 15.56%, respectively. Concerning Indiana dataset, the (OA and AA) were equal to (86.93% and 91.32%), (87.48% and 91.75%), (90.67% and 94.22%) and (91.05% and 94.81%) respectively. The gain in accuracy with respect to SVM were (5.77% and 4.16%), (6.32% and %4.59), (9.51% and 7.06%) and (9.89% and 7.65%), respectively. The worst class accuracy was obtained for the "ω_5 – *Soybeans-min till*" class and was

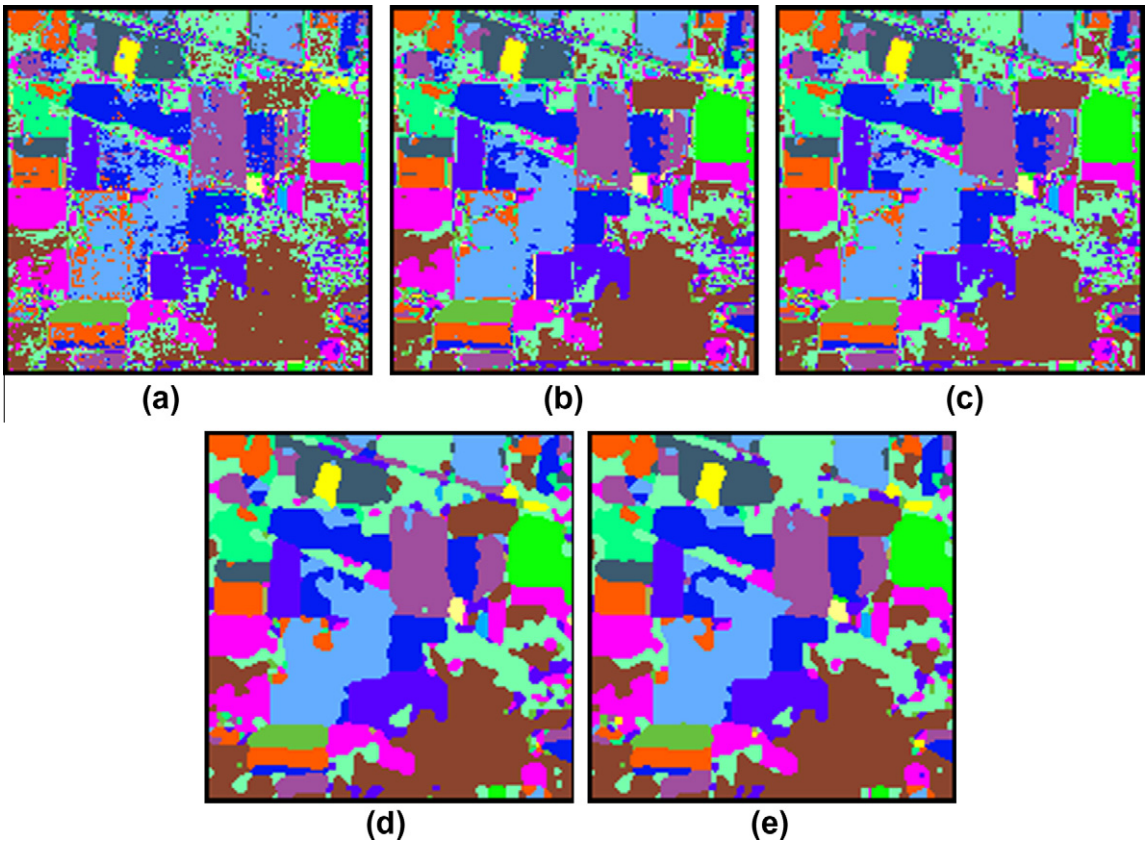

Fig. 7. Classification maps obtained by (a) SVM, (b) MV, (c) WMV, (d) MV-MRF and (e) WMV-MRF for the Indiana dataset.

77.81%, 80.10%, 84.44% and 85.73%, respectively. The gains with respect to SVM were equal 8.95%, 11.24%, 15.58% and 16.87%, respectively.

From these results, one can observe the following: (1) both pixel-based and MRF-based fusion methods provided significant results with respect to the state-of-the-art SVM classifier; (2) the MRF-based fusion methods provided better classification results compared to the pixel-based methods confirming the advantage of including contextual information in the decision process; (3) from a statistical point of view, the WMV-MRF method is the most promising fusion method followed by the MV-MRF method as reported in Table 4; and (4) from a qualitative point of view, Figs. 6 and 7 show that these methods provided us with more realistic classification maps.

4.2.5. Sensitivity with respect to the ensemble size P

In this experiment, we assessed the sensitivity of the four fusion methods with respect to the ensemble size P. We repeated the above experiments for different values of this parameter (i.e., 5, 20, 30, and 40). Figs. 8 and 9 show the trend of the (OA and AA) accuracies and the computation time, respectively, with respect to P. As can be seen, in general the accuracies were stable and exhibited an increasing trend with increasing values of P. For all cases, the results were significant in terms of classification accuracy with respect to SVM. For example, for Pavia dataset the (OA and AA) obtained by the MV, WMV, MV-MRF, and WMV-MRF fusion methods for $P = 5$ were equal to (92.46% and 93.66%), (92.22% and 93.58%), (95.01% and 95.99%), and (95.02% and 96.14%), respectively. By contrast, for $P = 40$ the (OA and AA) become (94.02% and 95.39%), (93.77% and 95.35%), (96.15% and 97.38%), and (96.16% and 97.53%), respectively. The time taken for generating the results was (173, 173, 996, and 552 s) and (1320, 1320, 2290 and 1830 s), respectively for both cases. Concerning Indiana dataset, the (OA and AA) yielded by the fusion methods for $P = 5$ were equal to (86.29% and 91.00%), (86.86% and 91.35%), (90.39% and 94.03%) and (90.80% and 94.66%), respectively. These accuracies become (87.13% and 91.38%), (87.70% and 91.76%), (90.85% and 94.29%) and (91.08% and 94.86%), respectively, for $P = 40$. The computation time for both cases was (43, 43, 165, and 98 s) and (134, 134, 296 and 220 s), respectively. From these results, one can again confirm the great capability of the proposed fusion methods in improving the classification accuracy with respect to SVM. In particular, the ability of the Markovian fusion methods in boosting further the results compared to those based on voting rules.

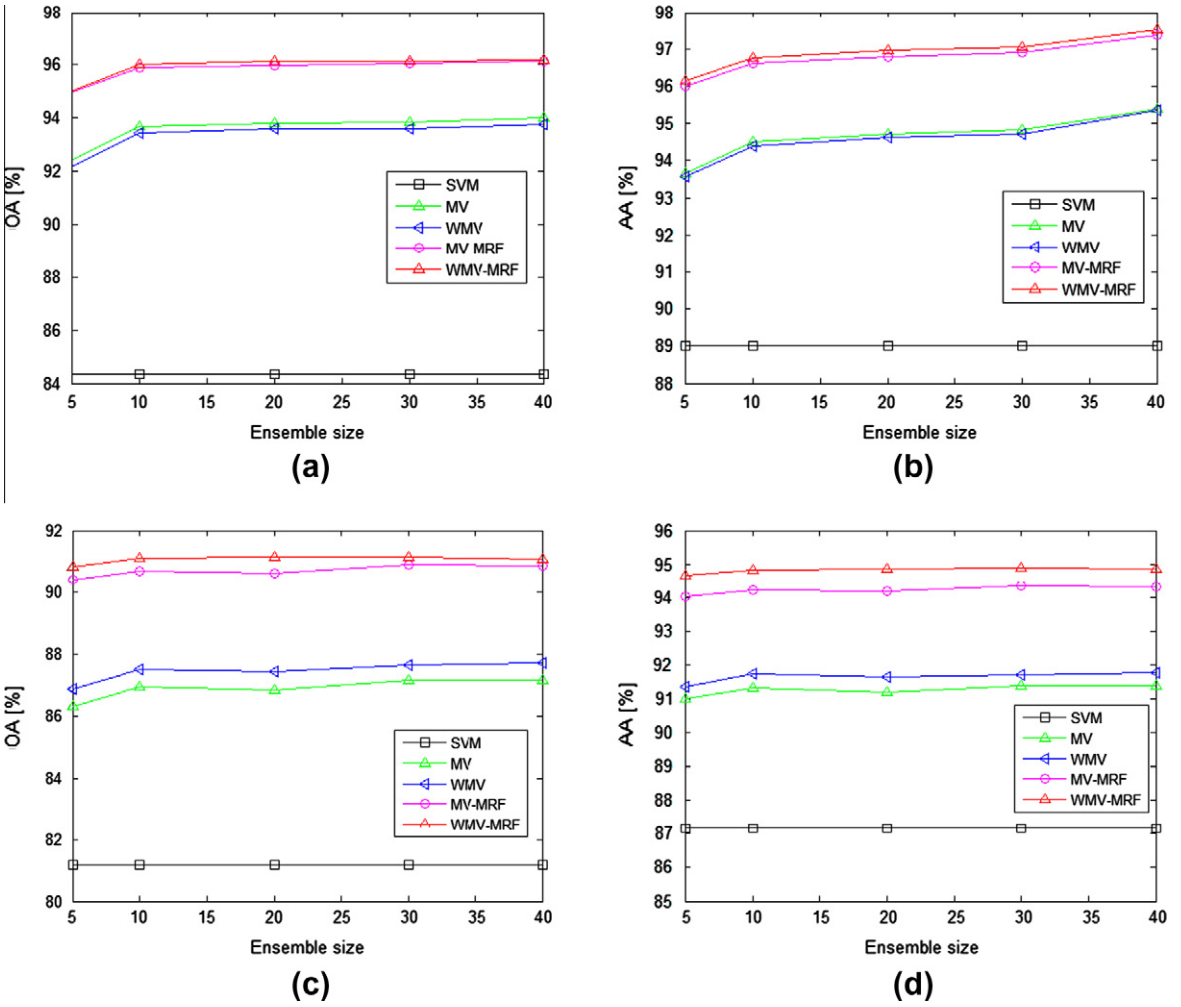

Fig. 8. OA and AA accuracies achieved by the proposed fusion methods versus the size of the ensemble P for (a)–(b) the Pavia and (c)–(d) the Indiana datasets.

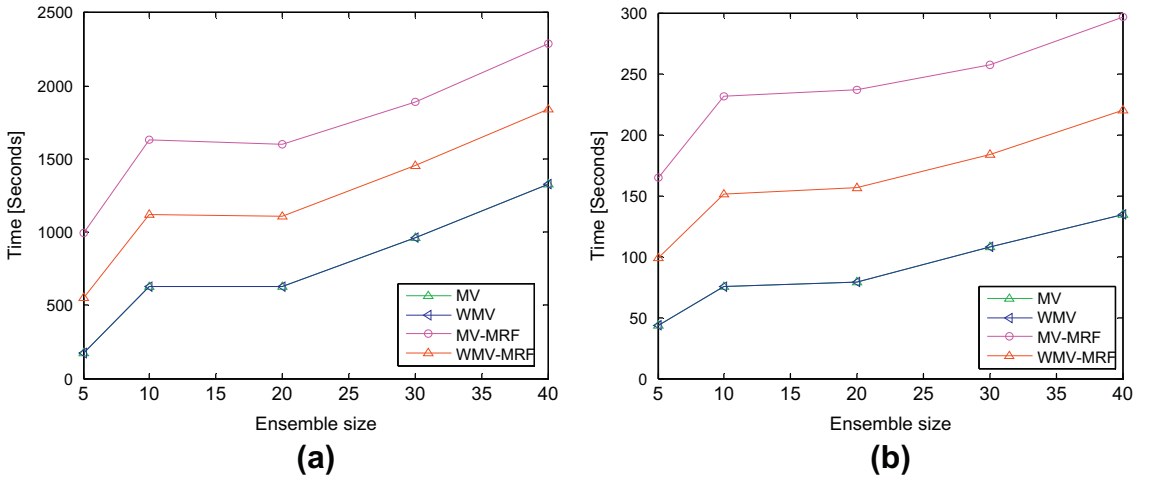

Fig. 9. Computation time taken by the proposed fusion methods versus the size of the ensemble P for (a) the Pavia and (b) the Indiana datasets.

5. Conclusions

This paper has proposed a novel framework for combining the results of supervised and unsupervised learning methods. Different fusion schemes based on simple voting rules, and Markovian theory were presented. The experimental results obtained on two hyperspectral images allow us to draw the following conclusions: (1) All fusion methods were able to provide significant improvements in classification with respect to the state-of-the-art SVM; (2) The MRF-based fusion methods performed the best compared to those based on voting rules as they exploit contextual information in decision making; and (3) The computation time is relatively contained. Future developments of this framework could be defined in two main directions. The adoption of more robust unsupervised learning methods for segmenting the hyperspectral images and the investigation of sophisticated fusion operators in order to boost further the classification accuracy.

Acknowledgement

The authors would like to acknowledge the support from the Distinguished Scientist Fellowship Program at King Saud University. The authors would like also to thank Prof. D. Landgrebe and Prof. P. Gamba (University of Pavia, Italy) for providing the hyperspectral images used in the experiments.

References

[1] A. Agresti, Categorical Data Analysis, second ed., Wiley, New York, 2002.
[2] K. L Bakos, P. Gamba, Hierarchical hybrid decision tree fusion of multiple hyperspectral data processing chains, IEEE Trans. Geosci. Remote Sens. 49 (2011) 388–394.
[3] Y. Bazi, F. Melgani, Toward an optimal SVM classification system for hyperspectral remote sensing images, IEEE Trans. Geosci. Remote Sens. 44 (2006) 3374–3385.
[4] Y. Bazi, F. Melgani, L. Bruzzone, G. Vernazza, A genetic expectation maximization method for unsupervised change detection in multitemporal SAR imagery, Int. J. Remote Sens. 30 (2009) 6591–6610.
[5] J.A. Benediktsson, J.R. Sveinsson, Multisource remote sensing data classification based on consensus and pruning, IEEE Trans. Geosci. Remote Sens. 41 (2003) 932–936.
[6] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York, 1981.
[7] S. Bourennane, C. Fossati, A. Caill, Improved of classification for hyperspectral images based on tensor modeling, IEEE Geosci. Remote Sens. Lett. 7 (2010) 801–805.
[8] L. Bruzzone, F. Melgani, Robust multiple estimator systems for the analysis of biophysical parameters from remotely sensed data, IEEE Trans. Geosci. Remote Sens. 46 (2005) 159–174.
[9] J.B. Campbell, R.H. Wynee, Introduction to Remote Sensing, fifth ed., The Guilford press, 2011.
[10] Q. Cheng, H. Zhou, J. Cheng, The Fisher-Markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 1217–1233.
[11] R.C. Dubes, A.K. Jain, Random field models in image analysis, J. Appl. Statist. 16 (1989) 131–163.
[12] M. Fauvel, J. Chanussot, J.A. Benediktsson, Decision fusion for the classification of urban remote sensing images, IEEE Trans. Geosci. Remote Sens. 44 (2006) 2828–2838.
[13] S. Fereidoun, A. Mianji, Y. Zhang, Robust hyperspectral classification using relevance vector machine, IEEE Trans. Geosci. Remote Sens. 49 (2011) 2100–2112.
[14] G.M. Foody, Thematic map comparison: evaluating the statistical significance of differences in classification accuracy, Photogram. Eng. Remote Sens. 70 (2004) 627–633.
[15] N. Ghoggali, F. Melgani, Y. Bazi, A Multiobjective genetic SVM approach for classification problems with limited training samples, IEEE Trans. Geosci. Remote Sens. 47 (2009) 1707–1718.
[16] A. Ghosh, N.S. Mishra, S. Ghosh, Fuzzy clustering algorithms for unsupervised change detection in remote sensing images, Inform. Sci. 181 (2011) 699–715.
[17] C.W. Hsu, C.J. Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Netw. 13 (2002) 415–425.
[18] G. Jun, J. Ghosh, Spatially adaptive classification of land cover with remote sensing data, IEEE Trans. Geosci. Remote Sens. 49 (2011) 2662–2673.
[19] H.R. Kalluri, S. Prasad, L.M. Bruce, Decision-level fusion of spectral reflectance and derivative information for robust hyperspectral land cover classification, IEEE Trans. Geosci. Remote Sens. 48 (2010) 4047–4058.
[20] X. Liu, Models to determine parameterized order weighted averaging operators using optimization criteria, Inform. Sci. 190 (2012) 27–55.
[21] J. Li, J.M. Bioucas-Dias, A. Plaza, Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields, IEEE Trans. Geosci. Remote Sens. 50 (2012) 809–823.
[22] C. Li, B. Kuo, C. Lin, C. Huang, A spatial contextual support vector machine for remotely sensed image classification, IEEE Trans. Geosci. Remote Sens. 50 (2012) 784–799.
[23] L. Ma, M.M. Crawford, J. Tian, Local manifold learning-based κ-nearest-neighbor for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 48 (2010) 4099–4109.
[24] S. Maldonado, R. Weber, J. Basak, Simultaneous feature selection and classification using kernel-penalized support vector machines, Inform. Sci. 181 (2011) 115–128.
[25] F. Melgani, L. Bruzzone, Classification of hyperspectral remote sensing images with support vector machine, IEEE Trans. Geosc. Remote Sens. 42 (2004) 1778–1790.
[26] F. Melgani, Y. Bazi, Markovian fusion approach to robust unsupervised change detection in remotely sensed imagery, IEEE Geosci. Remote Sens. Lett. 3 (2006) 457–461.
[27] F. Melgani, S.B. Serpico, A Markov random field approach to spatio-temporal contextual image classification, IEEE Trans. Geosci. Remote Sens. 41 (2003) 2478–2487.
[28] G. Mountrakis, J. Im, C. Ogole, Support vector machines in remote sensing: a review, ISPRS J. Photo. Remote Sens. 66 (2011) 247–259.
[29] M.D. Mura, A. Villa, J.A. Benediktsson, J. Chanussot, L. Bruzzone, Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis, IEEE Geosci. Remote Sens. Lett. 8 (2011) 542–546.
[30] M. Picco, G. Palacio, Unsupervised classification of SAR images using Markov random fields and $\mathcal{G}_I^0$ models, IEEE Geosci. Remote Sens. Lett. 8 (2011) 350–353.
[31] J. Platt, Probabilities for support vector machines, in: A. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (Eds.), Avances in Large Margin Classifiers, MIT Press, Cambridge, MA, 2000, pp. 61–74.

[32] S. Prasad, L.M. Bruce, Decision fusion with confidence based weight assignment for hyperspectral target recognition, IEEE Trans. Geosci. Remote Sens. 46 (2008) 1448–1456.
[33] G.X. Ritter, G. Urcid, a latex matrix method for hyperspectral image unmixing, Inform. Sci. 181 (2011) 1787–1803.
[34] S.B. Serpico, G. Moser, Weight parameter optimization by the Ho-Kasyap algorithm in MRF models for supervised image classification, IEEE Trans. Geosci. Remote Sens. 44 (2006) 3695–3705.
[35] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. Roy. Stat. Soc. B 36 (1974) 111–147.
[36] Y. Tarabalka, J.A. Benediktsson, J. Chanussot, Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques, IEEE Trans. Geosci. Remote Sens. 47 (2009) 2973–2987.
[37] Y. Tarabalka, J. Chanussot, J.A. Benediktsson, Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers, IEEE Trans. Syst. Man Cyber. Part B: Cyber. 40 (2010) 1267–1279.
[38] Y. Tarabalka, M. Fauvel, J. Chanussot, J.A. Benediktsson, SVM- and MRF- based method for accurate classification of hyperspectral images, IEEE Geosci. Remote Sens. Lett. 8 (2010) 736–740.
[39] T. Udelhoven, V. Linden, B. Waske, M. Stellmes, L. Hoffmann, Hypertemporal classification of large areas using decision fusion, IEEE Geosci. Remote Sens. Lett. 6 (2009) 592–596.
[40] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[41] B. Waske, S. Linden, J.A. Benediktsson, A. Rabe, P. Hostert, Sensitivity of support vector machines to random feature selection in classification of hyperspectral data, IEEE Trans. Geosci. Remote Sens. 48 (2010) 2880–2889.
[42] Z. Xu, R.R. Yager, Intuitionistic fuzzy Bonferroni means, IEEE Trans. Syst. Man Cybern. Part B Cyber. 41 (2011) 568–578.
[43] Z. Xu, R.R. Yager, Power geometric operators and their use in group decision making, IEEE Trans. Fuzzy Syst. 18 (2010) 94–105.
[44] R.R. Yager, On the fusion of imprecise uncertainty measures using belief structures, Inform. Sci. 181 (2011) 3199–3209.
[45] R.R. Yager, J. Kacprzyk, The Ordered Weighted Averaging Operators: Theory and Applications, Kluwer, Norwell, MA, 1997.
[46] R.R. Yager, On ordered weighted averaging aggregation operators in multi-criteria decision making, IEEE Trans. Syst. Man Cyber. 18 (1988) 183–190.
[47] R.R. Yager, Generalized OWA aggregation operators, Fuzzy Optim. Decis. Making 3 (2004) 93–107.
[48] R.R. Yager, On generalized Bonferroni mean operators for multicriteria aggregation, Int. J. Appro. Reason. 50 (2009) 1279–1286.
[49] L. Zhang, L. Zhang, D. Tao, X. Huang, On combining multiple features for hyperspectral remote sensing image classification, IEEE Trans. Geosci. Remote Sens. 50 (2012) 879–893.
[50] P. Zhong, R. Wang, Modeling and classifying hyperspectral imagery by CRFs with sparse higher order potentials, IEEE Trans. Geosci. Remote Sens. 49 (2011) 688–705.