Hyperspectral Image Classification with Limited Labeled Training Samples Using Enhanced Ensemble Learning and Conditional Random Fields

Fan Li, Linlin Xu, *Member, IEEE*, Parthipan Siva,
Alexander Wong, *Member, IEEE*, David A. Clausi, *Senior Member, IEEE*

***Abstract*—Classification of hyperspectral imagery using few labelled samples is a challenging problem considering the high dimensionality of hyperspectral imagery. Classifiers trained on limited samples with abundant spectral bands tend to overfit, leading to weak generalization capability. Therefore, it is crucial to develop classification approaches that are capable of reducing model variance, while at the same time learning general structure from the training samples in an efficient manner. To address this problem, we have developed an enhanced ensemble method called multiclass boosted rotation forest (MBRF), which combines the rotation forest algorithm and a multiclass AdaBoost algorithm. The benefit of the combination can be explained by bias-variance analysis, especially in the situation of inadequate training samples and high dimensionality. Further, MBRF innately produces posterior probabilities inherited from AdaBoost, which are served as the unary potentials of the conditional random field (CRF) model to incorporate spatial context information. Experimental results show that the classification accuracy of MBRF as well as its integration with CRF consistently outperforms the other referenced state-of-the-art classification methods when limited labeled samples are available for training.**

***Index Terms*—Hyperspectral data classification, rotation forests, AdaBoost, conditional random fields.**

I. Introduction

Classification of hyperspectral imagery is more difficult than other remote sensing imagery due to the Hughes phenomenon [1], also known as "curse of dimensionality". Given sufficient labeled training samples, the difference between different classifiers is negligible because they all converge at or close to the Bayes error rate. However, this is unrealistic in practice, especially for classification of remote sensing data for which the acquisition of ground truth is usually expensive and time-consuming. Due to the difficulty and costs of obtaining ground truth for remote sensing imagery, classification using few labeled samples, i.e., the "small-sample-size (SSS) classification" has attracted the attention of remote sensing researchers in recent years [2], [3]. This problem of inadequate training samples is deteriorated by the high dimensionality of spectral bands in hyperspectral images. One way to address this problem is to incorporate unlabeled samples using semi-supervised classification methods [4], [5]. However, previous empirical experiments show that it is possible that semi-supervised methods make no improvement or even have detrimental impact on classification performance [6]. Other ways to overcome the limitation of labeled samples include exploiting the spectral-spatial information of hyperspectral data by feature extraction and feature selection techniques [3] or leveraging its sparsity nature by sparse representation methods [7].

Ensemble methods have been successfully applied for hyperspectral image classification. Besides well-known methods such as random forests [8] and AdaBoost [9], more and more advanced ensemble methods have been recently proposed [10]–[13]. Compared to other classifiers, dimension reduction is usually unnecessary for ensemble methods because they deal fairly well with high-dimensional data [14]. In recent years, ensemble methods have been shown in particular to achieve high classification performance when the number of training samples is limited. Waske et al. [15] demonstrated that classifier ensembles using support vector machines and random feature selection can significantly improve classification performance. Yang et al. [2] proposed a dynamic subspace method for hyperspectral image classification which achieves better classification accuracy than the random subspace method. Xu et al. [16] investigated different classifiers for marine oil spill identification, and found that bagging-based methods significantly outperform other classification methods.

A recent trend of hyperspectral image classification research is the prevalence of spectral-spatial classification methods that incorporate spatial context to improve classification performance [1]. It has been demonstrated that global random field methods are better than local filtering methods [17]. The most commonly-used random field model for remote sensing imagery is Markov random field (MRF) [1]. Previous literature showed that the classification performance can be improved by combining MRF with a unary classifier that can generate probabilistic outputs [18]–[20]. In recent years, a powerful discriminative random field model, i.e., the conditional random field (CRF) model has been used for remote sensing imagery. Zhong and Wang [21] first introduced CRF to classification of hyperspectral images using multinomial logistic regression (MLR) as its unary classifier, and demonstrated its superiority to MRF. Zhang and Jia [22] modified the CRF model in [21] by incorporating a boundary constraint in order to reduce model parameters to train. They all demonstrated that the CRF models have better classification performance than the traditional MRF models.

This paper addresses the training sample inadequacy issue by proposing a novel spectral-spatial hyperspectral image classification approach based on an enhanced ensemble classifier and the conditional random field (CRF) technique. First, the proposed multiclass boosted rotation forest (MBRF)

algorithm that integrates rotation forest and AdaBoost is used to obtain pixelwise estimation. The motivation is based on bias-variance analysis, i.e., the two ensemble classifiers have relative advantages in terms of decreasing model bias and model variance. Therefore, the combination of the two, as conducted in the proposed enhanced ensemble classifier, is able to take advantage of merits of both classifiers, especially in this small training sample size context. Second, in order to model the spatial contextual information in hyperspectral image, the proposed algorithm is incorporated into the CRF framework, serving as the unary term in the CRF objective function. Experiments on benchmark hyperspectral images demonstrate that the proposed MBRF algorithm is capable of outperforming other referenced state-of-the-art classifiers, and is better able to aid the CRF approach for spectral-spatial classification of hyperspectral image, especially when the number of training samples is small.

The rest of this paper is organized as follows: Section II discusses the relative advantages and inadequacies of two mainstream ensemble methods, i.e., boosting and bagging, from a perspective of the trade-off issue between model bias and variance, which motivates the proposed enhanced ensemble method. Section III introduces the proposed framework, including the details of the MBRF algorithm, its ability to approximate posterior probability estimates, and its incorporation into CRF to achieve spectral-spatial classification. Section IV shows the comparison of the proposed method and several referenced state-of-the-art classification methods on three hyperspectral datasets. Section V concludes the paper by summarizing the above sections.

II. Review of Ensemble Methods

Ensemble methods are computational techniques that combine a large number of base classifiers for improved prediction [23]. They have been widely used for supervised classification due to their flexibility, ease of implementation, and outstanding performance.

The benefit of ensemble methods can be explained from a bias-variance decomposition perspective, which was originally proposed by Geman et al. for a regression model of squared loss [24]. Domingos [25] provided a unified bias-variance decomposition which can be applied to any loss function. The predicted loss $E_{D,y}[L(y,h)]$ for a given loss function L can be decomposed into intrinsic noise, bias and variance:

$$\begin{aligned} &E_{D,y}[L(y,h)] \\ &= c_1 E_y[L(y,h_*)] + L(h_*,h_m) + c_2 E_D[L(h_m,h)] \\ &= c_1 N(x) + B(x) + c_2 V(x) \end{aligned} \tag{1}$$

where D is the training set, x is the example, y is the true value, h is the prediction, h_* is the optimal prediction that minimizes $E_y[L(y,h_*)]$, h_m is the main prediction [25] which is the mean of the predictions under squared loss, c_1 and c_2 are constants, and $N(x)$, $B(x)$, and $V(x)$ represent noise, bias, and variance respectively.

Since the noise is irreducible, we only consider the bias and the variance. The bias describes the error of the classifier in expectation, and the variance reflects the sensitivity of the classifier to variations in the training samples. For squared loss, $c_1 = c_2 = 1$, so both bias and variance increase the predicted loss. However, for zero-one loss used in classification problems, it has been demonstrated that c_2 is negative for biased examples, and therefore there is a much higher tolerance for variance [20], [25].

Ideally, we wish the classifiers to have both low bias and low variance. However, there is usually a tradeoff between bias and variance [26]. When the complexity of a classifier goes up, the bias tends to decrease while the variance will increase. Ensemble methods can largely reduce the variance by majority voting of the results by base classifiers, without affecting the bias or event reducing it [27]. Therefore, weak learners that are sensitive to small changes in data are selected as base classifiers, such as decision trees and perceptron [26].

Two mainstream ensemble approaches are boosting [28] and bagging [23]. The idea of boosting methods is to learn classifiers iteratively by adjusting the distribution of training samples based on the classification error, and predict labels by weighted majority voting. Both bias and variance can be reduced by boosting. Empirical experiments show that boosting is not easy to overfit [29], but when the sample size is small, boosting tends to have large variance and thus cause overfitting [30]. One of the most popular boosting methods is the AdaBoost Algorithm [28]. It can be viewed as a forward stagewise additive modeling algorithm to minimize the exponential loss function.

The standard bagging aims to reduce model variance by performing majority voting among base classifiers that are trained on bootstrap subsets of training samples. Compared to boosting, the bagging technique can reduce more model variance, but the model bias is unchanged. In recent years, there are also some variants of the bagging method that have become popular. The random forest algorithm [31] is a method that combines bagging and random subspace method [32]. A reduction in the variance can be achieved by reducing the correlation between the trees, at the expense of a slight bias increase. Another typical method is the rotation forest algorithm [33], which performs feature transformation with some randomness on the data for each base classifier, and also combine the results by majority voting. Previous empirical experiments show that rotation forests have smallest variance compared to other ensemble methods such as boosting and random forests. Also, unlike bagging and random forests, the rotation can reduce bias even though the reduction of bias is not as significant as boosting [34].

Motivated by the above bias-variance analysis, an ensemble method integrating rotation forests and AdaBoost is proposed in this paper so that both bias and variance can be further reduced by taking advantage of both methods. A similar approach to our proposed method is the RotBoost algorithm [35], but it is not developed for small-sample-size context. Moreover, RotBoost adopts standard AdaBoost algorithm which requires the training error of the base classifiers to be less than $1/2$, which is too strict for a multiclass classification problem. Instead, we use here a multiclass AdaBoost algorithm that is more tolerant about the training error. We also show

that the proposed method allows the posterior probability to be naturally obtained without requiring the base classifiers to estimate probabilities. The posterior probability is used into the CRF framework to incorporate spatial context information. The implementation details will be introduced in the next section.

III. Proposed Framework

In this section, we propose a two-stage framework for hyperspectral image classification with limited number of training samples. The first stage is to perform ensemble learning to obtain posterior probability of class labels for pixels in hyperspectral image using only spectral information. The MBRF algorithm that combines rotation forests and AdaBoost is proposed for improving label prediction based on the motivation in Section II. Using a multiclass AdaBoost algorithm, the posterior probability can be naturally generated without requiring the base classifiers to output probability estimates. In the second stage, based on the posterior probability, the proposed MBRF classifier is incorporated into the CRF framework, in order to simultaneously incorporate both spectral and spatial information in hyperspectral imagery.

A. Multiclass Boosted Rotation Forest

The MBRF method is a bagging-based method that combines multiple classifiers which are independent from each other. Instead of performing bootstrap sampling, the data is first perturbed by performing feature transformation with randomness, and then it is trained by adaptive boosting. Therefore, the individual classifier for MBRF is not a single base classifier, but a boosted ensemble of base classifiers which is called meta-base classifier (MBC). Any rotation-variant classifier, i.e., the decision boundary will be changed by rotating the feature space of the data, can be used as base classifiers. The posterior probabilities by an MBC can be naturally approximated, and the probabilities by all the MBCs are finally combined. The flow chart of the MBRF algorithm is shown in Fig. 1.

To train a classifier h_i, the first step is to perturb the original data by multiplying by a rotation matrix. To increase the randomness of the rotation matrix, the original feature set is randomly divided into Q subsets, and a random number of classes are eliminated. Then a bootstrap sample of 75% sample size is selected [33]. Afterwards, a feature extraction method is performed on the bootstrap sample without reducing the dimensions. Empirical experiments show that principal component analysis (PCA) is the feature extraction method that can achieve best classification performance [10], [36]. The coefficients by PCA obtained for each subset are incorporated into a rotation matrix:

$$R = \begin{bmatrix} c_1^{(1)}, c_1^{(2)}, \cdots, c_1^{(M_1)} & \cdots & [0] \\ [0] & & \\ \vdots & \ddots & \vdots \\ [0] & \cdots & c_Q^{(1)}, c_Q^{(2)}, \cdots, c_Q^{(M_Q)} \end{bmatrix} \quad (2)$$

where Q is the number of subsets, M_i is the number of variables in each subset i ($i = 1...Q$), and $c_i^{(M_1)}, \cdots, c_i^{(M_i)}$ are $M_i \times 1$ coefficient vectors of principal components obtained from the bootstrap samples with variables in the i^{th} subset.

The columns in R are rearranged according to the order of the original feature set to obtain the final rotation matrix R^a. The procedure of calculating the rotation matrix is shown in Alg. 1.

Algorithm 1 Calculating rotation matrix R^a

1: Split the feature set $\mathbf{F}$ into Q subsets: F_i, $(i = 1, ..., Q)$;
2: **for** $i = 1...Q$ **do**
3: Let X_i be the dataset X for features in F_i;
4: Remove a random subset of classes from X_i;
5: Select a bootstrap sample of 75% sample size from X_i to form a new sample set X_i';
6: Apply PCA on X_i' to obtain the coefficients c_i^j, $(j = 1, ..., M_i)$;
7: **end for**
8: Construct R with the obtained coefficients using Eq. (2);
9: Output R^a by rearranging the columns of R by matching the order of features in F.

The second step is to perform AdaBoost on the rotated data. The original AdaBoost algorithm [28] is for binary classification. For multiclass problems, it may easily fail when the training error by base classifiers is greater than $1/2$. One way is to use one-versus-rest or one-versus-one strategies to decompose into multiple binary classification problems [37], such as AdaBoost.MH [38] and AdaBoost.M2 [39]. A disadvantage of these methods is the posterior probability cannot be directly generated. In the proposed method, we use a multiclass AdaBoost algorithm called SAMME [40]. It can be considered as forward stagewise additive modeling using a multiclass exponential loss function, which has the same statistical explanation as the original AdaBoost algorithm. Given K classes, it only requires the error rate by a base classifier to be less than $1/K$ rather than $1/2$ which is very rigid for multiclass classification. We can also see that SAMME will reduce to AdaBoost when $K = 2$.

At the beginning, the training data are sampled from the training set D in the uniform distribution, i.e.,$\mathfrak{D}_1(\mathbf{x}) = 1/n$, where n is the number of training samples. Then, a base classifier h_t is trained on the sampled data, and the error ϵ^t is calculated:

$$\epsilon^t = P_{\mathbf{x_t} \sim \mathfrak{D}_t}(h_t(\mathbf{x}) \neq y) \quad (3)$$

where $\mathbf{x}$ is the training samples and y is their true labels.

The sampling distribution is updated in the way that misclassified samples are assigned larger weights:

$$\mathfrak{D}_{t+1}(\mathbf{x}) = \frac{1}{Z_t}\mathfrak{D}_t(\mathbf{x}) \cdot exp\{\ \alpha^t \mathbb{I}(h_t(\mathbf{x}) \neq y)\ \} \quad (4)$$

where Z_t is a normalization factor, $\mathbb{I}(\cdot)$ is the indicator function, and α_t is the model parameter based on the training error:

$$\alpha^t = log\frac{1-\epsilon^t}{\epsilon^t} + log(K-1) \quad (5)$$

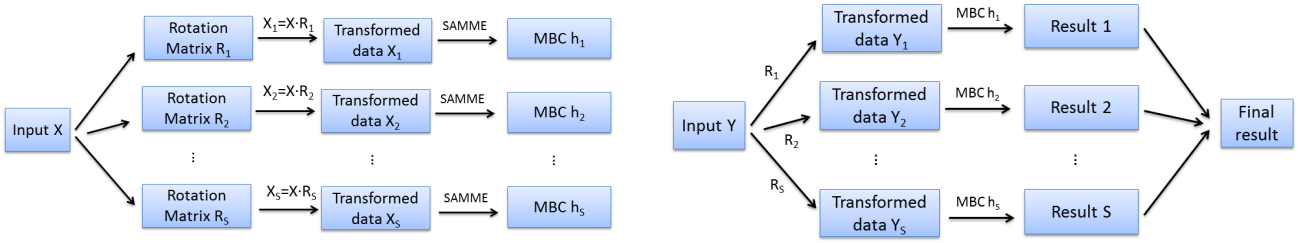


Fig. 1. Flow charts of the proposed MBRF algorithm. The left is training stage and the right is test stage. S rotation matrices and S classifiers are learned based on training data X in the training stage. They are used to estimate the posterior probabilities of test data Y in the test stage.

For a new sample $\mathbf{x}$, the posterior probability $P(y = k \mid x)$ for each MBC can be approximated:

$$P(y = k \mid x) = \frac{e^{\frac{1}{K-1}} f_k^*(x)}{e^{\frac{1}{K-1}} f_1^*(x) + \cdots + e^{\frac{1}{K-1}} f_K^*(x)} \tag{6}$$

where

$$f_k^*(x) = \sum_{t=1}^{T} \alpha_t \cdot \delta(h_t(x), k) \tag{7}$$

and

$$\delta(h_t(x), k) = \begin{cases} 1 & h_t(x) = k \\ -\frac{1}{K-1} & \text{otherwise} \end{cases} \tag{8}$$

After the posterior probabilities for each MBC are obtained by SAMME, the final probability estimates are calculated by averaging over all the boosted classifiers. The MBRF algorithm is described in Algorithm 2. Eq. (6) provides a way to obtain class probability with good theoretical meaning, so that it is not necessary for a single base classifier to generate probability estimates.

Algorithm 2 The MBRF algorithm

1: **for** $s = 1$ to S: **do**
2: Calculate the rotation matrix R_s^a as shown in Alg. 1.
3: Perturb the data by multiplying the rotation matrix: $\mathbf{x_s} = \mathbf{x} \cdot R_s^a$;
4: Initialize the weight distribution $\mathfrak{D}_{s1}(\mathbf{x_s}) = 1/n$, $i = 1, 2, ..., n$.
5: **for** $t = 1$ to T: **do**
6: Train a classifier h_{st} from the training set D under distribution $\mathfrak{D}_{st}$: $h_{st} = \mathfrak{L}(D, \mathfrak{D}_{st})$;
7: Update the sampling distribution $\mathfrak{D}_{s,t+1}$ using Eq. (4);
8: **end for**
9: Output conditional probability $P_s(y = k \mid \mathbf{x_s})$ using Eq. (6).
10: **end for**
11: Output final probability estimates by averaging: $P(y = k \mid \mathbf{x}) = \frac{1}{S} \sum_{s=1}^{S} P_s(y = k \mid \mathbf{x_s})$.

B. Conditional random fields

The traditional MRF model is formulated in a probabilistic generative framework modeling the joint probability of the image and its labels [41], [42]. It assumes that a set of random variables have a Markovian property, which means the random variables are only dependent on their neighborhood. According to Bayes rule, the posterior probability is modeled as $P(\mathbf{y} \mid \mathbf{x}) \propto P(\mathbf{x} \mid \mathbf{y})P(\mathbf{y})$, where $\mathbf{y}$ is the labels and $\mathbf{x}$ is the observations. $P(\mathbf{y})$ is modeled as Gibbs distribution. $P(\mathbf{x} \mid \mathbf{y})$ can be represented as a factorized form if we assume the conditional probability $P(x_i \mid y_i)$ is independent:

$$P(\mathbf{x} \mid \mathbf{y}) = \prod_i P(x_i \mid y_i) \tag{9}$$

Contrary to MRF, CRF [43] discriminatively models the posterior probability directly so that the rigid conditional independence assumption can be relaxed:

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} exp\left\{ -\sum_{c \in C} \psi_c(\mathbf{y_c}, \mathbf{x}) \right\} \tag{10}$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} exp\{ -\sum_{c \in C} \psi_c(\mathbf{y_c}, \mathbf{x}) \}$ is the partition function and ψ_c is a potential defined on clique c.

For simplification, only unary and pairwise clique potentials are usually considered. Eq. (10) can be rewritten as

$$\begin{aligned} P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} exp\{ -[& \sum_{i \in S} \phi_i(y_i, \mathbf{x}) \\ & + \sum_{i \in S} \sum_{j \in \eta_i} \xi_{ij}(y_i, y_j, \mathbf{x})] \} \end{aligned} \tag{11}$$

where $\phi_i(\cdot)$ and $\xi_{ij}(\cdot)$ are unary and pairwise clique potentials respectively, η_i is the set of neighbors of site i, and S is the sample set.

In this paper, an 8-connected CRF is used. The unary and pairwise clique potentials can be defined as arbitrary domain-specific local discriminative classifiers [21]. In previous remote sensing literature, the unary potentials have been defined as posterior probabilities by various discriminative classifiers such as multinomial logistic regression [22], [44] and support vector machines [18], [45]. In this paper, we use the probability estimates by the proposed MBRF method as shown in the previous section. Thus, the unary potential can be defined as

$$\phi_i(y_i, \mathbf{x}) = \sum_{k=1}^{K} \delta(y_i = k) \{- \, logP(y_i = k \mid \mathbf{x_i}) \} \quad (12)$$

where $P(y_i = k \mid \mathbf{x_i})$ is calculated using Eq. (6).

For the pairwise potentials, the standard MRF model only allows the contextual information of the labels to be used (i.e., the standard Potts model [46]), while both the labels and the observed data can be formulated in the CRF model as $\xi_{ij}(y_i, y_j, \mathbf{x})$ in Eq. (11). We note that pairwise connected terms will tend to have different labels at discontinuities in image structure. As a result, we use a generalized Potts model as the discontinuity preserving smoothness constraint:

$$\xi_{ij}(y_i, y_j, \mathbf{x}) = \begin{cases} \beta exp\{-\alpha(E_i + E_j)/2\} & y_i \neq y_j \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

where E is the edge image obtained using the Gaussian derivative per band then per pixel maximum over all the bands, $\alpha = 1/(0.25 \cdot T_{otsu})$, T_{otsu} is the Otsu Threshold of the edge image E which will adapt the edge strength based on global edge strength of the image [47], and β is a constant representing the degree of smoothness.

In practice, the optimum β is usually selected using cross validation [45]. In the inference step, the optimal labeling is assigned by maximizing the posterior probability in Eq. (11), i.e., to solve the energy minimization problem below:

$$argmin_y \sum_{i \in S} \phi_i(y_i, \mathbf{x}) + \sum_{i \in S} \sum_{j \in \eta_i} \xi_{ij}(y_i, y_j, \mathbf{x}) \quad (14)$$

In previous literature, loopy belief propagation (LBP) [48] is usually used to solve this combinatorial optimization problem [21], [45], but recently graph-cut based methods have become popular. For binary labeling problem, graph-cuts method can find the global optimum [49]. Boykov et al. [50] developed an efficient graph-cut based algorithm, i.e., α-expansion and α-β-swap algorithms, which are able to find approximate solution to the multiclass labeling problem. The α-expansion algorithm has been demonstrated to outperform other state-of-the-art energy minimization methods on benchmark problems [51]. Furthermore, it has been proved [50] that a local minimum within a known factor of the global minimum can be found using α-expansion. As a result we use the α-expansion algorithm to solve the energy minimization problem.

IV. Experimental Results and Analysis

In this section, experiments are conducted for testing the performance of the proposed MBRF method and the combination with CRF. First, MBRF is compared with several state-of-the-art pixelwise classification methods, including support vector machines (SVM), random forests (RF), SAMME, and rotation forests (RoF). SVM is one of the most commonly-used classifier for remote sensing imagery. RF and SAMME are advanced versions of Bagging and standard AdaBoost respectively. Previous literature showed that feature extraction can improve the classification performance of RF [20]. Therefore, we also test the performance of RF after using minimum noise fraction (MNF) [52], which is a noise-adjusted version of PCA. Then, the combination of MBRF and CRF without using edge strength (MBRFCRF-NE) and with edge strength (MBRFCRF-E) is tested with two recently proposed spectral-spatial methods: SVMMRF-E [18] and MLRCRF-E [22]. It is noted that SVMMRF-E is actually a CRF-based method because it uses posterior probabilities by a discriminative classifier and uses edge strength in the pairwise potentials.

We use three standard hyperspectral datasets: the Universtiy of Pavia, the Indian Pine, and the Kennedy Space Center datasets for testing in our experiments. The main objectives of this section are testing and evaluating the performance of the aforementioned pixelwise and spectral-spatial classification methods in different numbers of training samples for the three datasets.

A. Experimental Setup

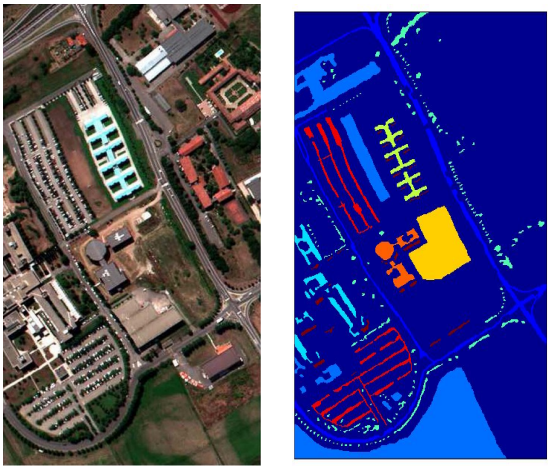
For SVM, the radial basis function (RBF) is used as the kernel function, and the optimum parameters, i.e., the regularization parameter C and the bandwidth parameter of the RBF kernel γ are found using 5-fold cross-validation. The number of trees is set to 500. The number of variables randomly selected at each split is set by default, i.e., the square root of the total number of variables. For MNF, we use all the projected features whose performance has been tested to be better than that using a subset. For SAMME, the number of base classifiers are set to 100. For RoF and MBRF, the number of classes eliminated from the original data to calculate the rotation matrix is fixed to 3. In previous literature, the number of trees in the rotation forest algorithm is usually set to 10 [10], [33]. In our paper, the small-sample-size problem might leads to slow convergence, so the number of trees in rotation forests is set to 50. For MBRF, the number of MBCs is set to 30, and the number of trees in an MBC is set to 20. Decision tree classifier is used as the base classifier for all the ensemble methods. Due to the incapability of generating posterior probabilities directly by SVM, pairwise coupling [53] used in [18] is also adopted in our experiment.

We use the same setting for the number of training samples as [3]. Different numbers of randomly selected training samples per class (3, 5, 10, and 15) are used for testing. The overall accuracy (OA), average accuracy (AA), and Kappa statistic are calculated to evaluate the classification performance. We also notice that the classification performance is very sensitive to the selection of training samples when the number of training samples is limited, so all the methods are tested for 50 times using different randomly selected training samples, and the mean is used for all the statistics. For the smoothness parameter β in the CRF model, the traditional cross-validation is usually incapable of selecting the optimal parameter because the number of training samples is limited. Therefore, we conduct multiple tests using different β $(2^0, 2^1, 2^2, ..., 2^8)$ and report the highest test accuracy. To investigate the usefulness of the edge penalty information used in the smoothness constraint, we make a comparative study between the CRF

model using edge penalty in Eq. 13 and without using edge penalty (i.e., $\alpha \to +\infty$).

B. Experiments with the University of Pavia dataset

The first dataset for testing was acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor in University of Pavia, Italy (Fig. 2). The original image has 115 spectral bands with a spectral range from $0.43\mu m$ to $0.86\mu m$. 103 spectral reflectance bands are used for the analysis after the noisy bands are removed. The spatial resolution of the image is 1.3 meters. The size of the dataset used in the experiment is 610×340 pixels. There are nine classes in total.



(a) Data (b) Ground truth

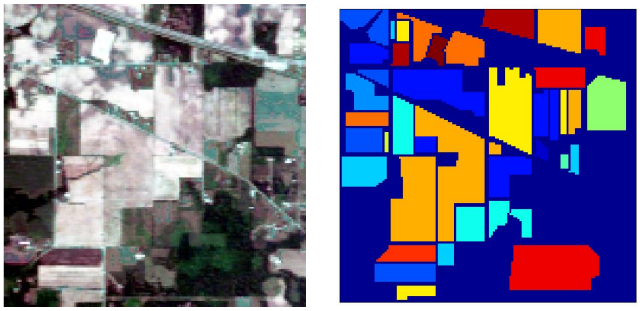
Fig. 2. The false-color composition of the University of Pavia data and its ground truth

The classification result of the University of Pavia dataset is shown in Table I. It is observed that MBRF achieves best classification performance for different number of labeled training samples. When there are only three samples per class, MBRF generally achieves about 5.5% OA higher than SVM, 9.2% higher than RF, 13.6% higher than MNF-RF, 16.0% higher than SAMME, and 2.4% higher than RoF. The classification accuracy increases more than ten percent on average by combining with CRF. The classification accuracy of each class in the case of 10 training samples per class is shown in Table II. MBRF has highest classification accuracy in five out of nine classes and has the highest average accuracy.

Among the spectral-spatial methods, MBRFCRF-E achieves highest classification accuracy in all the cases. Also, we can see that CRF with edge penalty is significantly better than that without edge penalty, with an improvement of 5.3% classification accuracy on average. Fig. 3 shows the segmentation result using MBRF and CRF with different number of training samples in one test. The object boundaries are well-delineated due to the high spatial resolution and the urban scene of the dataset, which can help extract a clean edge map for the CRF model. Using edge penalty, neighboring pixels with strong edges can be prevented to be assigned the same label. Meanwhile, higher smoothness parameter can be selected so that pixels with high within-class variation can be smoothed out in the labeling.

C. Experiments with the Indian Pine dataset

The Indian Pine dataset as shown in Fig. 4 was acquired from the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor from Northwest Indiana rural area on June 12, 1992 [54]. It has 224 spectral reflectance bands ranging from $0.4\mu m$ to $2.5\mu m$. The spatial resolution is 20 meters. The dataset used in this experiment is in the size of 145×145 pixels which is extracted from a larger scene. Also, 24 bands covering the water absorption bands are removed [54]. The number of classes is 16.



(a) Data (b) Ground truth

Fig. 4. False-color composition of the Indian Pine data and its ground truth

The classification result of the Indian Pine dataset is shown in Table III. The overall accuracy achieved by proposed method is 9.5% higher than SVM, 11.4% higher than RF, 6.9% higher than MNF-PCA, 15.7% higher than SAMME, and 2.5% higher than RoF on average.

The pixelwise classification map using 10 samples per class in one test is shown in Fig. 5. Due to the high within-class variation nature of the Indian Pine dataset, it is not enough to achieve satisfactory classification result by using only spectral information. In this case, the CRF model which serves as a smooth labeling method can help improve classification significantly. Both MBRFCRF-E and MBRFCRF-NE increase OA by about 14 percent in all the cases compared to the pixelwise MBRF method. Also, they both achieve more than five percent higher OA than SVMMRF-E and MLRCRF-E. However, it seems that edge penalty does not help improve the classification performance in this dataset. Compared with the above ROSIS dataset, this dataset has relatively low spatial resolution and high within-class variation, so the extract gradient map is noisy and does not reflect the true edges in the scene.

D. Experiments with the Kennedy Space Center dataset

The Kennedy Space Center (KSC) dataset as shown in Fig. 6 was acquired by the AVIRIS instrument over the Kennedy Space Center, Florida on March 23, 1996 [8]. The original image has 224 bands with a spectral range from $0.4\mu m$ to $2.5\mu m$. 176 bands are used in the experiment after the water absorption and noisy bands are removed. The spatial resolution of the image is 18 meters. The size of the dataset used in the experiment is 512×614 pixels. There are 13 land cover types in the training area of the reference data.

TABLE I
OVERALL ACCURACY (OA%) AND AVERAGE ACCURACY (AA%) , AND KAPPA COEFFICIENT ([κ]) OF DIFFERENT NUMBER OF TRAINING SAMPLES BY DIFFERENT METHODS FOR THE UNIVERSITY OF PAVIA DATASET.

Classifier	Labeled samples per class			
	3	5	10	15
SVM	56.6 (68.0) [0.47]	60.0 (71.4) [0.51]	68.9 (77.3) [0.61]	73.8 (80.7) [0.67]
RF	56.7 (68.1) [0.47]	59.8 (70.6) [0.51]	62.4 (73.7) [0.54]	65.5 (75.6) [0.57]
MNF-RF	39.5 (54.1) [0.29]	54.6 (66.9) [0.45]	63.5 (74.6) [0.55]	69.1 (77.9) [0.61]
SAMME	42.6 (51.4) [0.32]	53.2 (62.9) [0.43]	58.6 (69.3) [0.49]	62.8 (72.7) [0.54]
RoF	59.2 (67.9) [0.50]	64.1 (73.9) [0.56]	71.9 (79.3) [0.65]	76.5 (82.5) [0.70]
MBRF	61.6 (71.9) [0.53]	67.6 (76.2) [0.60]	73.7 (80.7) [0.67]	78.3 (84.0) [0.72]
SVMMRF-E	52.2 (61.3) [0.43]	66.0 (71.5) [0.58]	83.9 (86.1) [0.79]	90.8 (92.4) [0.88]
MLRCRF-E	71.2 (76.2) [0.64]	79.6 (82.6) [0.74]	88.1 (88.5) [0.85]	91.7 (92.3) [0.89]
MBRFCRF-NE	69.1 (75.9) [0.61]	78.2 (82.7) [0.72]	85.6 (88.6) [0.82]	90.2 (92.5) [0.87]
MBRFCRF-E	75.2 (80.3) [0.69]	83.7 (86.9) [0.79]	90.9 (92.9) [0.88]	94.6 (95.9) [0.93]

TABLE II
CLASS ACCURACY ACHIEVED BY DIFFERENT PIXELWISE CLASSIFIERS FOR THE UNIVERSITY OF PAVIA DATASET IN THE CASE OF 10 TRAINING SAMPLES PER CLASS.

Classifier	C1	C2	C3	C4	C5	C6	C7	C8	C9	AA
SVM	65.9	64.3	58.5	90.5	99.1	58.4	85.8	73.2	**99.9**	77.3
RF	65.0	53.4	52.3	87.1	99.0	50.1	82.7	70.6	**99.9**	73.4
MNF-RF	51.9	58.9	**69.9**	**92.2**	**99.9**	62.0	83.9	52.1	99.8	74.5
SAMME	62.2	49.7	50.0	81.2	97.1	52.6	77.1	66.4	91.4	69.8
RoF	67.6	67.3	61.5	91.0	99.0	66.9	83.8	**76.0**	99.8	79.2
MBRF	**70.4**	**70.0**	67.4	91.9	99.4	**68.7**	**87.2**	75.5	**99.9**	**81.2**

TABLE III
OVERALL ACCURACY (OA%) AND AVERAGE ACCURACY (AA%), AND KAPPA COEFFICIENT ([κ]) OF DIFFERENT NUMBER OF TRAINING SAMPLES BY DIFFERENT METHODS FOR THE INDIAN PINE DATASET.

Classifier	Labeled samples per class			
	3	5	10	15
SVM	40.5 (52.1) [0.34]	46.2 (58.7) [0.40]	54.7 (67.3) [0.49]	60.5 (71.1) [0.56]
RF	41.1 (52.4) [0.34]	44.9 (57.2) [0.39]	51.9 (63.9) [0.46]	56.6 (67.0) [0.51]
MNF-RF	37.5 (51.1) [0.31]	50.3 (64.4) [0.45]	60.3 (74.3) [0.56]	64.4 (76.9) [0.60]
SAMME	32.3 (41.4) [0.26]	40.6 (52.3) [0.34]	49.5 (61.9) [0.44]	54.8 (65.7) [0.49]
RoF	44.5 (56.8) [0.39]	53.1 (65.7) [0.48]	63.5 (74.8) [0.59]	68.9 (78.2) [0.65]
MBRF	47.6 (58.9) [0.42]	56.0 (67.6) [0.51]	65.8 (76.4) [0.62]	70.6 (79.3) [0.67]
SVMMRF-E	28.5 (38.6) [0.22]	42.9 (55.8) [0.37]	68.9 (77.7) [0.65]	77.2 (83.8) [0.74]
MLRCRF-E	55.9 (65.3) [0.51]	64.1 (72.1) [0.60]	74.3 (81.9) [0.71]	77.2 (84.2) [0.74]
MBRFCRF-NE	61.1 (67.5) [0.56]	69.9 (78.3) [0.66]	79.8 (87.7) [0.77]	83.4 (90.1) [0.81]
MBRFCRF-E	62.6 (70.1) [0.58]	70.7 (77.4) [0.67]	80.2 (86.4) [0.77]	83.1 (88.6) [0.81]

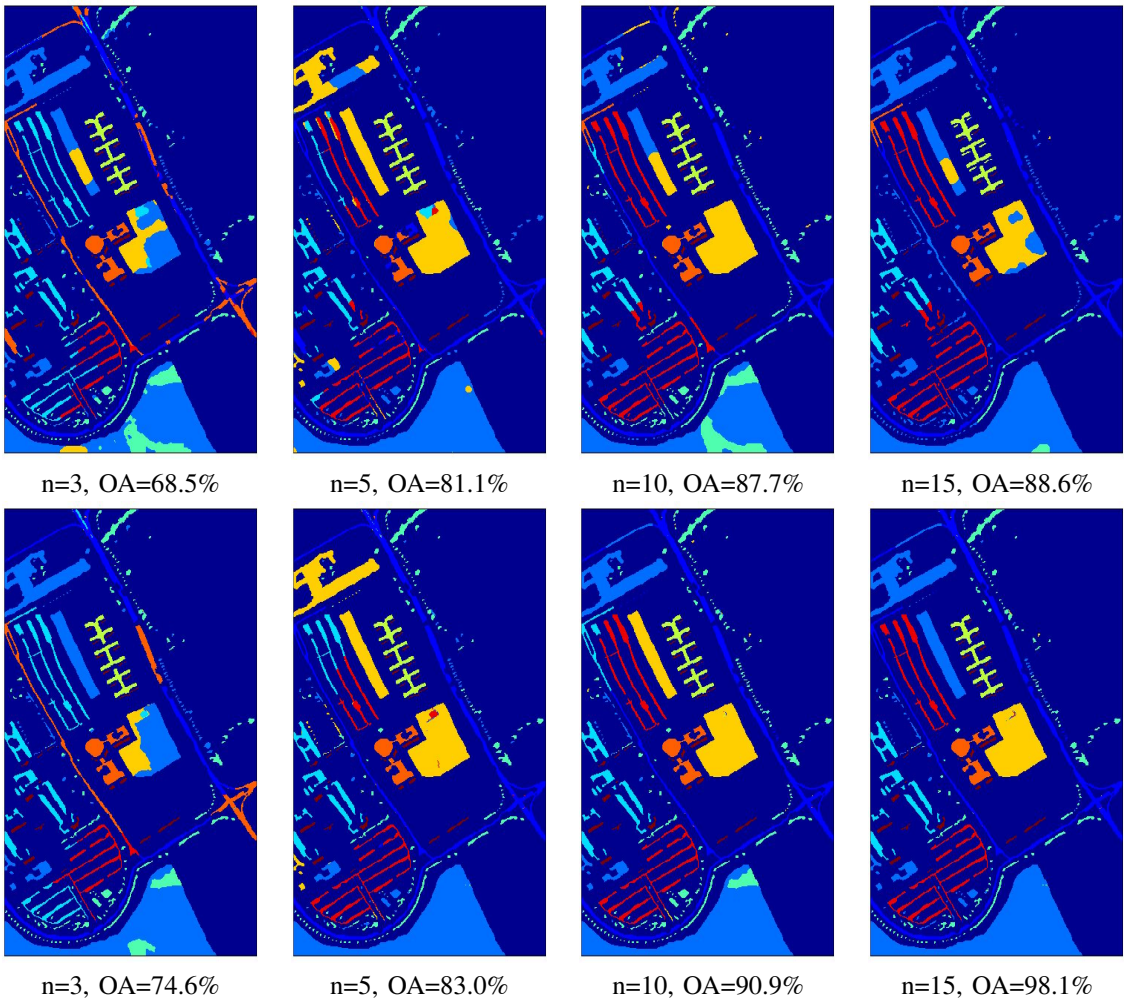


Fig. 3. Segmentation map by MBRFCRF-NE (top) and MBRFCRF-E (bottom) in different number of training samples (n per class) with optimal smoothness parameter in the University of Pavia dataset. Results in the same column are based on the same probability outputs by MBRF.

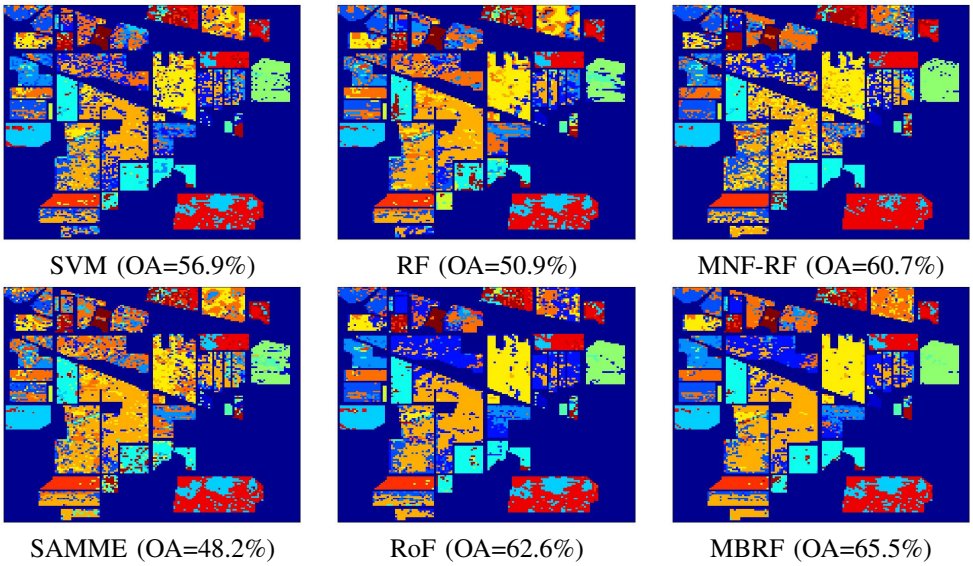
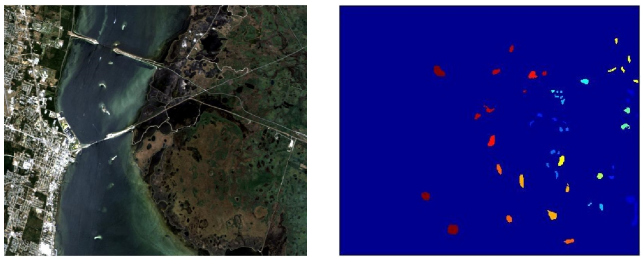


Fig. 5. Pixelwise classification results using 10 labeled training samples per class on the Indian Pine dataset with overall accuracy (OA%) of test samples in one test.

TABLE IV
OVERALL ACCURACY (OA%) AND AVERAGE ACCURACY (AA%) IN PERCENTAGE, AND KAPPA COEFFICIENT ([κ]) OF DIFFERENT NUMBER OF TRAINING SAMPLES BY DIFFERENT METHODS FOR THE KENNEDY SPACE CENTER DATASET.

Classifier	Labeled samples per class			
	3	5	10	15
SVM	73.1 (67.5) [0.70]	79.1 (74.2) [0.77]	86.0 (82.2) [0.84]	88.4 (85.0) [0.87]
RF	68.5 (62.2) [0.65]	73.5 (68.0) [0.70]	78.1 (73.6) [0.76]	80.9 (76.5) [0.79]
MNF-RF	65.8 (60.3) [0.62]	79.4 (74.1) [0.77]	86.0 (81.5) [0.84]	88.5 (84.5) [0.87]
SAMME	56.0 (51.2) [0.51]	69.2 (64.5) [0.66]	77.9 (73.7) [0.75]	81.9 (77.8) [0.80]
RoF	73.2 (67.6) [0.70]	80.0 (75.3) [0.78]	85.8 (81.8) [0.84]	88.2 (84.4) [0.87]
MBRF	78.0 (72.7) [0.76]	82.7 (78.2) [0.81]	87.5 (83.6) [0.86]	89.8 (86.2) [0.89]
SVMMRF-E	68.1 (62.0) [0.65]	82.8 (78.8) [0.81]	92.7 (91.0) [0.92]	95.2 (94.2) [0.95]
MLRCRF-E	84.3 (81.3) [0.82]	89.2 (87.0) [0.88]	94.4 (93.0) [0.94]	96.3 (95.2) [0.96]
MBRFCRF-NE	87.8 (84.5) [0.86]	91.9 (89.3) [0.91]	96.2 (94.7) [0.96]	98.0 (96.7) [0.98]
MBRFCRF-E	86.8 (84.4) [0.85]	90.9 (88.4) [0.90]	95.3 (94.0) [0.95]	97.3 (96.2) [0.97]



(a) Data (b) Ground truth

Fig. 6. The false-color composition of the Kennedy Space Center data and its ground truth

The classification result is shown in Table IV. Compared to the two datasets, this dataset is relatively easy because the pixelwise classification can achieve over 90% accuracy by MBRF using only 15 training samples for each class. Considering the classification accuracy for all the cases, MBRF averagely outperforms SVM by 2.8%, RF by 9.2%, MNF-RF by 4.6%, SAMME by 13.2%, and RoF by 2.7%.

After combining with CRF, the highest overall accuracy using 15 training samples per class can reach 98% by MBRFCRF-NE. There is no improvement of using the edge penalty in the CRF model. Similar to the Indian Pine dataset, this dataset has relatively low spatial resolution. Also, there is not much strong edge information in the rural area of the image where most of the training areas are located. As shown in Fig. 7, both methods achieve similar classification accuracy inside the training areas. However MBRFCRF-NE tends to oversmooth over the image, while MBRFCRF-E is better at preserving details, especially in small regions with strong boundary information.

E. Summary of Classification Results and Sensitivity Analysis

Observed from the pixelwise classification results (Table III, Table I, and Table IV), the overall classification accuracy by SAMME is not satisfactory because it is prone to overfitting when there are limited training samples. RF is less prone to overfitting because it reduces model variance, however, it cannot reduce bias of the decision tree classifiers. SVM is slightly better than RF, especially in the cases of $n = 10$ and $n = 15$. However, SVM is sensitive to the selection of model parameters, which are often determined by cross validation. When the number of training samples is limited, the parameters that achieve best cross-validation performance are less likely to be the optimal parameters. This could be improved using unsupervised heuristics or semi-supervised heuristics. Using an additional feature extraction step, MNF-RF achieves better performance than RF in 8 out of 12 cases, especially when the training samples are more than ten per class. However, its performance is still no better than RoF that has an inherent PCA transformation step. The rotation forest classifier which reduces both bias and variance perform well very in small-sample-size problems. It achieves higher overall classification accuracy than other comparing methods except MBRF in most of the cases. MBRF achieves the highest OA, AA, and Kappa for all the cases, and it gains additional improvement of 2.5% classification accuracy over RoF. This is because the combination of RoF with SAMME can further reduce model bias. Moreover, the high variance and overfitting drawback of boosting in small-sample-size problem turns into a benefit because it can de-correlate the MBCs from each other, and thus increase diversity.

From the spectral-spatial classification results, we observe that OA achieved by SVMMRF-E is not satisfactory in the cases of $n = 5$ and $n = 10$, which is even worse than the

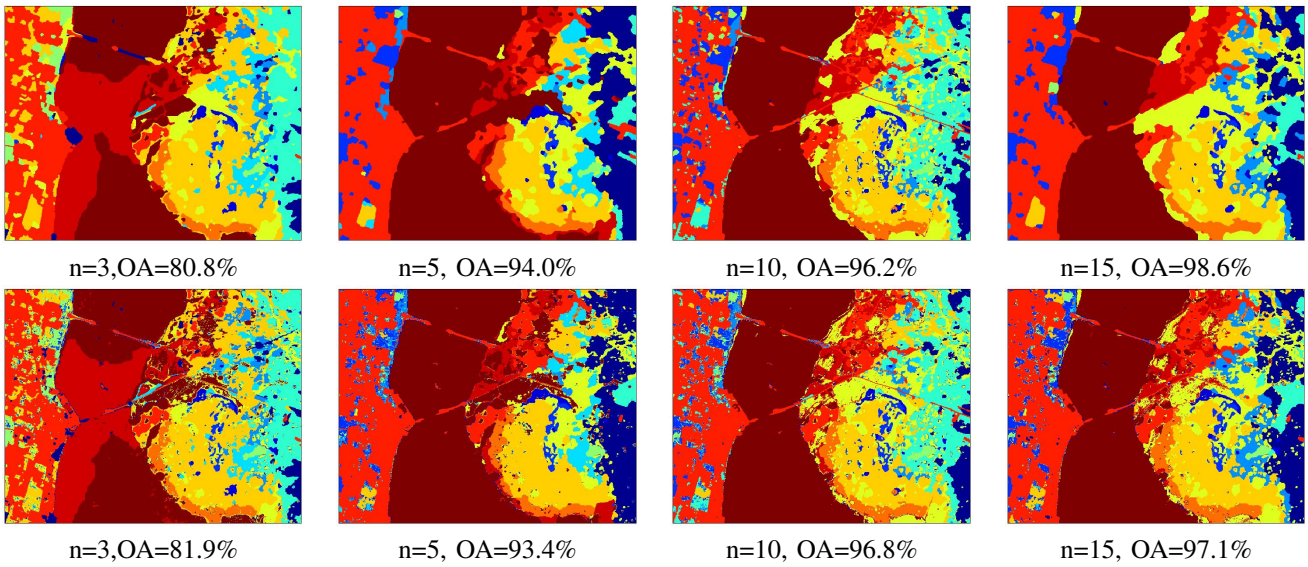


Fig. 7. Segmentation map by MBRFCRF-NE (top) and MBRFCRF-E (bottom), in different number of training samples (n per class) with optimal smoothness parameter β in the Kennedy Space Center dataset. We can see that details are more preserved in the results using edge penalty. Results in the same column are based on the same probability outputs by MBRF.

pixelwise SVM result. The reason might be that the pairwise coupling method [53] fails to generate satisfactory posterior probabilities when there are insufficient training samples. Instead, MLR is a classifier that learns posterior probabilities discriminatively, and results show that the OA achieves by MLRCRF-E is achieved better than that by SVMMRF-E. Similarly, MBRF allows the natural approximation of posteriors. We can see that the highest classification accuracy is achieved by either MBRFCRF-E or MBRFCRF-NE in all the datasets. Also, MBRFCRF-E achieves higher classification methods than MBRFCRF-NE in the University of Pavia dataset which has high spatial resolution and strong edge strength.

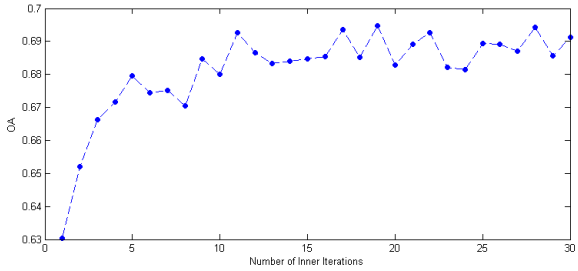


Fig. 8. Overall classification accuracy (OA%) as a function of the number of inner iterations T in MBRF for the University of Pavia dataset using different randomly-selected training samples (5 per class).

Compared to the rotation forest classifier, there is one more parameter to determine in the MBRF method, i.e., the number of trees T for the embedded SAMME algorithm. When $T = 1$, MBRF is reduced to RoF. Fig. 8 shows the overall classification accuracy as the number of trees increases for the University of Pavia dataset. The number of outer iterations S is fixed to 20, the same as the previous experimental setting. As shown in Fig. 8, the classification performance is improved as T increases, and the accuracy becomes stable when T is greater than 15.

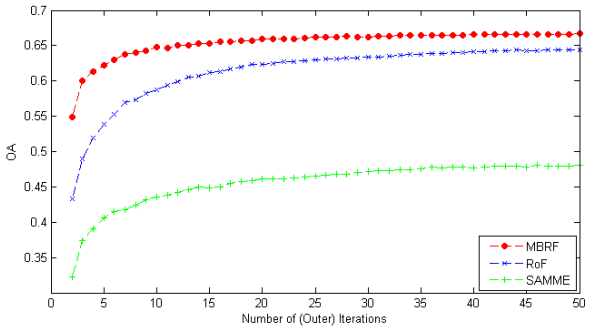


Fig. 9. Overall classification accuracy (OA%) as a function of the number of (outer) iterations S for the Indian Pine dataset using different randomly-selected training samples (10 per class).

We also test the convergence rate of MBRF by comparing with the rotation forest classifier. Fig. 9 shows the overall classification accuracy as the number of trees increases. For MBRF, the number of inner iterations T is set to 20. It is observed that MBRF converges after about 20 iterations, while SAMME and RoF converge after about 40 iterations. For the computation time, MBRF requires to train T base classifiers in an outer iteration compared to RoF, but the computational cost is still low considering the small training sample size, and the testing speed is very fast if the decision tree classifier is used.

Compared to other ensemble methods such as bagging and boosting, MBRF might be slower due to the innate PCA step. The computational complexity of PCA is $O(D^3)$ if the eigendecomposition of the $D \times D$ covariance matrix is performed using a power method [55]. But in MBRF, D is only a small subset of features, which is set to three in our experiment. Therefore, the PCA does not increase computation cost very much.

Finally, we test the sensitivity of the smoothness parameter.

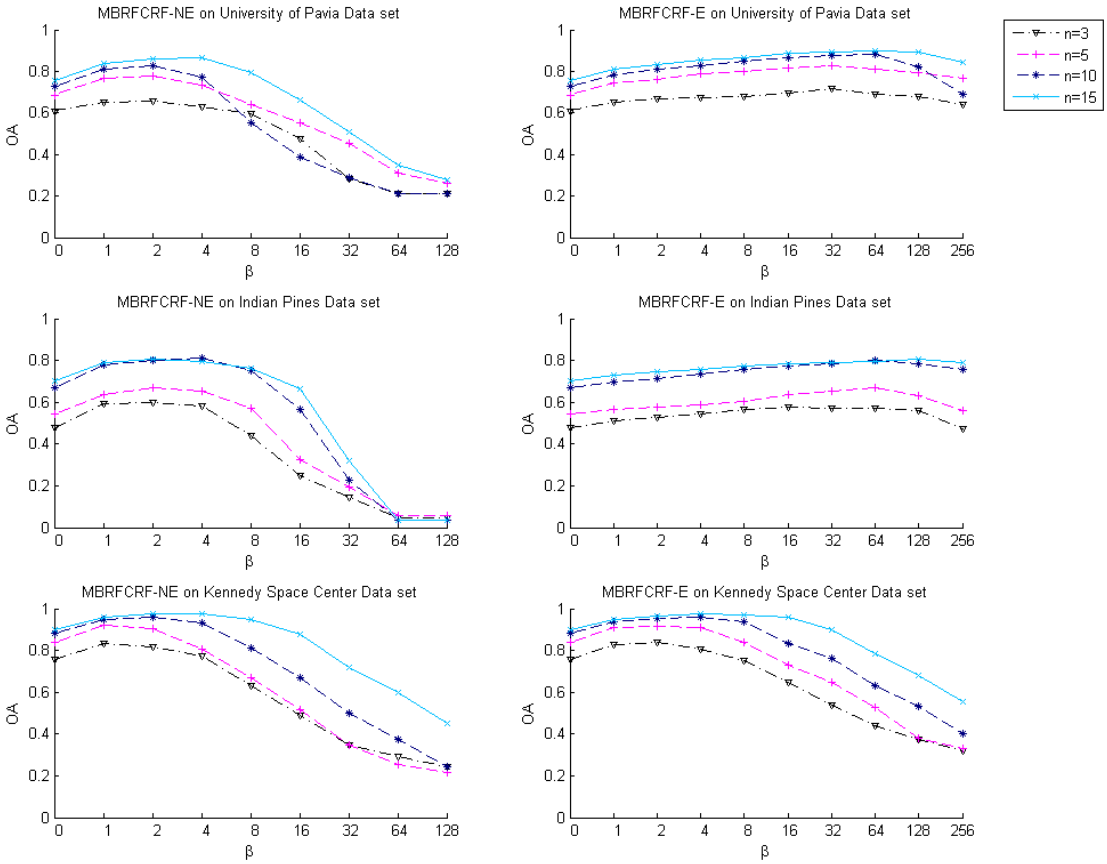


Fig. 10. Overall classification accuracy (OA%) for different smoothness parameter (β) and different number of training samples (n per class) in three datasets. The results in the left column are created by CRF without edge penalty, and the results in the right column are created by CRF with edge penalty.

The test accuracy related to different number of training samples for all the datasets is shown in Fig. 10. It is observed that the smoothness parameter for CRF with edge penalty is less sensitive than that without edge penalty. Also, the optimal smoothness parameter varies for different number of training samples. Based on the observations, it slightly increases when there are more training samples. One explanation is that the spatial context is only helpful when the probability estimates are reliable. If a majority of pixels in a region are misclassified, random fields or any other kind of smooth labeling methods tend to make the result worse.

V. Conclusions

A spectral-spatial classification method was proposed in this paper to deal with the situation when there are limited labeled training samples available. It is based on a novel ensemble method combining rotation forests and multiclass AdaBoost. The classification performance can be enhanced by combining both methods because the rotation forest algorithm reduces model variance while the AdaBoost algorithm reduces model bias. Also, we showed that the posterior probability can be naturally approximated by the proposed MBRF method and incorporated into the conditional random field framework. Experimental results showed that MBRF as well as its combination with CRF outperforms other state-of-the-art classification methods when the number of labeled training samples is limited.

VI. Acknowledgement

This work is partly funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Space Agency (CSA). The research was undertaken, in part, thanks to funding from the Canada Research Chairs program. The authors would like to thank Prof. P. Gamba, Prof. D. A. Landgrebe, and Prof. M. Crawford for sharing the hyperspectral datasets.

References

[1] A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri *et al.*, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, vol. 113, pp. S110–S122, 2009.

[2] J.-M. Yang, B.-C. Kuo, P.-T. Yu, and C.-H. Chuang, "A dynamic subspace method for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 7, pp. 2840–2853, 2010.

[3] S. Jia, Z. Zhu, L. Shen, and Q. Li, "A two-stage feature selection framework for hyperspectral image classification using few labeled samples," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1023–1035, 2014.

[4] M. Chi and L. Bruzzone, "Semisupervised classification of hyperspectral images by svms optimized in the primal," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1870–1880, 2007.

[5] G. Camps-Valls, T. Bandos Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, 2007.

[6] X. Zhu, "Semi-supervised learning literature survey," *Computer Science, University of Wisconsin-Madison*, vol. 2, p. 3, 2006.

[7] Q. Sami ul Haq, L. Tao, F. Sun, and S. Yang, "A fast and robust sparse approach for hyperspectral data classification using a few labeled samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2287–2302, 2012.

[8] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, 2005.

[9] S. Kawaguchi and R. Nishii, "Hyperspectral image classification by bootstrap adaboost with random decision stumps," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 11, pp. 3845–3851, 2007.

[10] J. Xia, P. Du, X. He, and J. Chanussot, "Hyperspectral remote sensing image classification based on rotation forest," *IEEE Geosci. Remote Sens. Lett.*, pp. 1–5, 2013.

[11] Y. Chen, X. Zhao, and Z. Lin, "Optimizing subspace svm ensemble for hyperspectral imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1306–1313, 2014.

[12] S. L. J. L. Alim Samat, Peijun Du and L. Cheng, "E^2LMs: Ensemble extreme learning machines for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1060–1069, 2014.

[13] P. Ramzi, F. Samadzadegan, and P. Reinartz, "Classification of hyperspectral data using an adaboostsvm technique applied on band clusters," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2066–2079, 2014.

[14] P. Yang, Y. Hwa Yang, B. B Zhou, and A. Y Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, no. 4, pp. 296–308, 2010.

[15] B. Waske, S. van der Linden, J. A. Benediktsson, A. Rabe, and P. Hostert, "Sensitivity of support vector machines to random feature selection in classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 7, pp. 2880–2889, 2010.

[16] L. Xu, J. Li, and A. Brenning, "A comparative study of different classification techniques for marine oil spill identification using radarsat-1 imagery," *Remote Sensing of Environment*, vol. 141, pp. 14–23, 2014.

[17] K. Schindler, "An overview and comparison of smooth labeling methods for land-cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4534–4545, 2012.

[18] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, "SVM- and MRF-based method for accurate classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, 2010.

[19] W. Li, S. Prasad, and J. E. Fowler, "Hyperspectral image classification using gaussian mixture models and markov random fields," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 153–157, 2014.

[20] A. Merentitis, C. Debes, and R. Heremans, "Ensemble learning in hyperspectral image classification: Toward selecting a favorable bias-variance tradeoff," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1089–1102, 2014.

[21] P. Zhong and R. Wang, "Learning conditional random fields for classification of hyperspectral images," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1890–1907, 2010.

[22] G. Zhang and X. Jia, "Simplified conditional random fields with class boundary constraint for spectral-spatial based remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 5, pp. 856–860, 2012.

[23] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[24] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992.

[25] P. Domingos, "A unified bias-variance decomposition," in *Proceedings of International Conference on Machine Learning*, 2000, pp. 231–238.

[26] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning*. Springer, 2009, vol. 2, no. 1.

[27] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *Proceedings of International Conference on Machine Learning*, vol. 96, 1996, pp. 148–156.

[28] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational Learning Theory*. Springer, 1995, pp. 23–37.

[29] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *The Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[30] R. Meir and G. Rätsch, "An introduction to boosting and leveraging," in *Advanced Lectures on Machine Learning*. Springer, 2003, pp. 118–183.

[31] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[32] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, 1998.

[33] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1619–1630, 2006.

[34] J. J. Rodríguez, C. J. Alonso, and O. J. Prieto, "Bias and variance of rotation-based ensembles," in *Computational Intelligence and Bioinspired Systems*. Springer, 2005, pp. 779–786.

[35] C.-X. Zhang and J.-S. Zhang, "Rotboost: A technique for combining rotation forest and Adaboost," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1524–1536, 2008.

[36] L. I. Kuncheva and J. J. Rodríguez, "An experimental study on rotation forest ensembles," in *Multiple Classifier Systems*. Springer, 2007, pp. 459–468.

[37] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC Press, 2012.

[38] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999.

[39] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Annals of Statistics*, pp. 1651–1686, 1998.

[40] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class Adaboost," *Statistics and Its*, 2009.

[41] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 6, pp. 721–741, 1984.

[42] S. Z. Li and S. Singh, *Markov random field modeling in image analysis*. Springer, 2009, vol. 26.

[43] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of International Conference on Machine Learning*, 2001, pp. 282–289.

[44] P. Zhong and R. Wang, "A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3978–3988, 2007.

[45] Y. Zhong, X. Lin, and L. Zhang, "A support vector conditional random fields classifier with a Mahalanobis distance boundary constraint for high spatial resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1314–1330, 2014.

[46] R. B. Potts, "Some generalized order-disorder transformations," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 48, no. 01. Cambridge Univ Press, 1952, pp. 106–109.

[47] P. Siva and A. Wong, "URC: Unsupervised regional clustering of remote sensing imagery," in *Proceedings of International Geoscience and Remote Sensing Symposium*, 2014.

[48] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *Proceedings of UAI*. Morgan Kaufmann Publishers Inc., 1999, pp. 467–475.

[49] D. Greig, B. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 271–279, 1989.

[50] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[51] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068–1080, 2008.

[52] A. A. Green, M. Berman, P. Switzer, and M. D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 65–74, 1988.

[53] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *The Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[54] D. A. Landgrebe, *Signal theory methods in multispectral remote sensing*. John Wiley & Sons, 2005, vol. 29.

[55] L. J. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research*, vol. 10, no. 1-41, pp. 66–71, 2009.