Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

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

Spectral–spatial hyperspectral image ensemble classification via joint sparse representation

Erlei Zhang [a], Xiangrong Zhang [a,b,*], Licheng Jiao [a], Lin Li [c], Biao Hou [a]

[a] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an 710071, China
[b] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[c] Key Laboratory of Information Fusion Technology, Ministry of Education, Northwestern Polytechnical University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 17 July 2015
Received in revised form
28 January 2016
Accepted 31 January 2016

Keywords:
Classification
Ensemble learning
Hyperspectral imagery
Joint sparse recovery
Spatial correlation

ABSTRACT

Ensemble learning can improve the performance of classification by integrating a set of classifiers, and shows significant potential benefits to the classification of hyperspectral image. However, the ensemble strategy remarkably influences the classification results, which include determining the minimum number of classifiers and assigning advisable weights associated with each classifier. In this paper, we present a novel sparse ensemble learning method with spectral–spatial knowledge for hyperspectral image classification. It considers the ensemble strategy under sparse recovery framework, where the solved non-zero coefficients reveal the importance of the selected classifier, from which a compact and effective ensemble learning system can be derived. Moreover, the spatial information is incorporated into the classification to develop a spectral–spatial joint sparse representation based ensemble learning algorithm for more accurate classification of hyperspectral images. Experimental results on several real hyperspectral images show that the proposed sparse ensemble system can achieve better performance than traditional ensemble learning methods using all classifiers, and it largely improves the efficiency in testing phase.

1. Introduction

Hyperspectral images (HSI) are generated from airborne and satellite sensors and contain abundant spectral information spanning the visible to the infrared spectrum. Nowadays, HSI has been fully applied to various fields, such as military [1–3], agriculture [4,5], and mineralogy [6]. HSI classification process is initiated to convert data into meaningful information where pixels are labeled to a certain class. Thus, classification of HSI is a very important task and research direction.

In pattern recognition, the class of a given pattern is usually decided by a single classifier. However, in remote-sensing information processing, it may be difficult for individual classifiers to make their own decisions in some cases. Here are some reasons as follows. Firstly, previous studies have found that no single classifier can be perceived as a "panacea" [7]. Each type of classification method has its own advantages and disadvantages for some certain problems. For example, a SVM classifier cannot make the best performance for all the scenarios of remote-sensing images. Secondly, complex data structures and relationships cannot be modeled by a convenient model. For instance, a pixel in HSI scene is generally a mixture of different materials with various abundance fractions, and its spectral information varies by atmospheric effects during data acquisition. Those variations are caused by uncertainty or randomness. This situation challenges the traditional single classifier based methods. Thirdly, some methods reach an improved performance by an ingenious "designing" effort, i.e. selection of parameters. However, it requires costly designing phase to obtain further improvements.

In order to solve those problems, nowadays ensemble learning has caused widespread interests [7–14]. The basic idea is to build an improved predictive model using some rules to integrate the various learning outcomes, thereby to obtain a better learning performance than a single learner. In a word, ensemble learning is a "many against one" problem. Ensemble classification has great potential in hyperspectral classification. Firstly, ensemble classification aims to take an effective combination method that makes use of each classifier but avoids the weakness. Secondly, in order to describe the characteristics of the data for classification, we can produce a set of "local specialized" classifiers for various feature space or overlapping regions. Combining the decision of "local experts", we can get satisfactory results. Finally, it can

* Correspondence to: Xidian University, P.O. Box 224, 710071 Xi'an, China. Tel.: +86 29 88202279; fax: +86 29 88201023.
E-mail addresses: zhangerlei0123@163.com (E. Zhang), xrzhang@ieee.org (X. Zhang), lchjiao@mail.xidian.edu.cn (L. Jiao), xdlinli86@163.com (L. Li), avcodec@163.com (B. Hou).

http://dx.doi.org/10.1016/j.patcog.2016.01.033
0031-3203/

reduce the computational burden in the designing phase through combining different classifiers. A lot of literatures have pointed out that the ensemble learning methods have a good effect on hyperspectral classification [12–22]. Some ensemble learning methods based on support vector machines (SVM) classifier have achieved good performance on hyperspectral classification [12–14]. Adaboost is a popular method to generate new training sets for ensemble learning with remote sensing data [15,16]. In most cases, it is used to improve the performance of "weak" learners such as decision trees [17]. The classification accuracy of Adaboost with decision trees is comparable to SVM [15]. But Adaboost is susceptible to noise and more costly in processing time [8]. Random Forest is another popular and effective method for classification of hyperspectral image [18–20]. It generates multiple trees from random subspaces of input features, then combining the resulting outputs via voting rule. Breiman [17] has proved that Random Forest has comparable performance with Adaboost. However, oversize ensemble system is not a good choice, because it costs too much. As well as both theoretical and empirical evidence [23,24] suggest that it is not always better for bigger size of ensembles. On the contrary, the small-sized ensembles can often reach a better performance than a larger one. Thus, the selection of classifiers and size of the ensemble system become problems need to be solved.

Recently, with the development of the compressed sensing theory [25,26], sparse representation has been introduced into computer vision and pattern recognition areas [27,28], as well as sparse modeling of signals and images [29–36]. Motivated by sparse representation, sparse ensembles as a new way of selective ensembles have caused extensive concern [37–40]. In sparse ensemble that generates a sparse weight vector, the nonzero coefficients correspond to the selected classifiers. However, those methods mainly focus on ℓ_1 norm constraint for obtaining sparse weights. Li et al. [41] proposed the idea to implement sparse ensemble through sparse reconstruction method, while the model handles sample independently without taking into account the data correlation. Furthermore, there are very few sparse ensemble methods in a particular field, for example, hyperspectral image.

In this paper, we consider ensemble learning problems as joint sparse reconstruction problems. We obtain sparse weights of classifiers from sparse recovery process, taking the characteristics of hyperspectral images into consideration simultaneously. Once obtaining large scale outputs of individual classifiers, the proposed method combines the classification outputs by a sparse recovery process. Sparse recovery process generates sparse solutions which can be considered as the weights of the individual classifier of ensemble system. This procedure, on one hand, selects a small part of individual classifiers to compose an ensemble system with a good classification performance, which dramatically reduces computational burden in the prediction process of new data sets; on the other hand, provides appropriate weights for the selected classifiers according to their relative importance. Moreover, contextual neighborhood knowledge can greatly improve the performance of the HSI classification [42–44]. So we formulate the ensemble learning with neighborhood sharing information into joint sparse reconstruction problems. Each pixel in the neighborhood presents one model and their information can be shared during the optimization process. The final ensemble solution is the fusion results of the pixels in the neighborhood. The experimental results show that the new ensemble system can automatically select lesser classifiers and perform better than traditional ensemble classification.

The paper is organized structures as follows. The related works are introduced in Section 2. And, the details of the proposed joint sparse representation-based ensemble learning method are shown in Section 3. Then, the effectiveness of the proposed method is demonstrated in Section 4. Finally, a summarization and some closing remarks are made in Section 5.

2. Related work

2.1. The general ensemble system framework

Algorithm 1 reveals the general procedure of an ensemble system. Suppose we get the training dataset $\{\boldsymbol{x}_i\}_{i=1,2,\cdots,N}$ and the corresponding true label $\{y_i\}_{i=1,2,\cdots,N}$. There are a variety of ways [49], including bagging, AdaBoost, random subspace, rotation forest, etc., to obtain a series of subsets of a training dataset for training different individual classifiers. The final classification map $\boldsymbol{y}^*$ is generated by combining the outputs $\boldsymbol{y}_i (i=1,\cdots,M)$ of all the individual classifiers with the voting rules, as summarized in Algorithm 1.

Algorithm 1. Generation of ensemble system

Input: Given training dataset $\boldsymbol{X}_{train}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]\in\Re^{d\times N}$, true label of the training dataset $\boldsymbol{y}_{train}=[y_1,y_2,\cdots,y_N]$, testing dataset $\boldsymbol{X}_{test}=[\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{x}}_2,\cdots,\tilde{\boldsymbol{x}}_{\tilde{N}}]\in\Re^{d\times\tilde{N}}$.

Output: Desired M classifiers for the ensemble and classification results $\boldsymbol{y}_{test}$ of the testing dataset.

Step 1: Use the training dataset to construct M classifiers $C_i(\boldsymbol{x})$, $i=1,2,\cdots,M$.

For $i=1:M$

(1) Generate a subset $(\boldsymbol{X}_i^{train},\boldsymbol{y}_i^{train})$of the training examples $(\boldsymbol{X}_{train},\boldsymbol{y}_{train})$;

(2) Generate a classifier, $C_i(\boldsymbol{x})$, trained on the subset $(\boldsymbol{X}_i^{train},\boldsymbol{y}_i^{train})$;

(3) Generate the predicted labels $\boldsymbol{y}_i^{test}$, for testing dataset $\boldsymbol{X}_{test}$, predicted by $C_i(\boldsymbol{x})$;

End for

Step 2: Generate the final classification results $\boldsymbol{y}_{test}$ by different fusion methods based on voting rule.

2.2. Voting rules

In the process of using M classifiers to solve the K-class classification problem, there are several common decision-making models based on an integrated approach of hard output.

(1) Majority voting (MV)

A vector $[c_{i,1},\cdots,c_{i,K}], c_{i,j}\in\{0,1\}$ can be obtained from classifier $C_i, i=1,\cdots,M$. Only one element in the vector is 1, the rest are 0. Sample belongs to jth class if $c_{i,j}=1$. The MV rule is described as follows:

$$y^*=\underset{j}{\operatorname{argmax}}\,\Upsilon_j, j=1,\cdots,K \qquad \Upsilon_j=\sum_{i=1}^{M}c_{i,j} \tag{1}$$

(2) Weighted majority voting (WMV)

$$\Phi(Aw)=\underset{j}{\operatorname{argmax}}\,\Upsilon_j, j=1,\cdots,K \quad \Upsilon_j=\sum_{i=1}^{M}w_{i,j} \tag{2}$$

Assume $OA_i^{tr}(i=1,\cdots,M)$ is overall accuracy on the training set of the ith classifier map, the weights are computed as follows:

$$w_i=\frac{OA_i^{tr}}{\sum_{j=1}^{M}OA_j^{tr}},\ \text{for } i=1,2,\cdots,M \tag{3}$$

Please cite this article as: E. Zhang, et al., Spectral–spatial hyperspectral image ensemble classification via joint sparse representation, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.033

The weights can be also computed as formula (4), which utilizes a priori knowledge of each classifier. Thus, the integrated system can maximize the accuracy of identification [47].

$$w_i = \frac{\log OA_i^{tr}/(1-OA_i^{tr})}{\sum_{j=1}^{M} \log\ OA_j^{tr}/(1-OA_j^{tr})}, \quad \text{for} \quad i = 1, 2, \cdots, M \tag{4}$$

From the perspective of the weighted majority voting, w_i is equal to 1 in MV rule. In the following part, making decisions based on MV rule means that all classifiers have equal weight. Here, the weights of classifiers equal to 1 or computed as (3) and (4) are denoted as MV, WMV1, and WMV2, respectively.

2.3. Sparse representation model

In the sparse representation model [27], a query signal can be approximated by a given dictionary under sparse coding framework. The key issue in this the reconstruction process is to find the sparse representation coefficient $\boldsymbol{\alpha} \in \Re^N$ for a testing sample $\boldsymbol{x} \in \Re^d$. Suppose a given dictionary $\boldsymbol{A} \in \Re^{d \times N}$ (e.g., consisted of all the training samples), the representation $\boldsymbol{\alpha}$ satisfying $A\boldsymbol{\alpha} = \boldsymbol{x}$ can be obtained as follows:

$$\hat{\boldsymbol{\alpha}} = \arg\min \left|\left|\boldsymbol{\alpha}\right|\right|_0 \quad s.t. \quad A\boldsymbol{\alpha} = \boldsymbol{x} \tag{5}$$

Due to the solution is difficult to approximate, the above problem can also be relaxed to a linear programming problem by ℓ_1 norm constraint [48] as follows:

$$\hat{\boldsymbol{\alpha}} = \arg\min ||\boldsymbol{\alpha}||_1 \quad \text{s.t.}\ ||A\boldsymbol{\alpha} - \boldsymbol{x}||_2 \le \varepsilon \tag{6}$$

where ε is the error tolerance.

3. Joint sparse representation-based ensemble algorithm

3.1. The proposed model

Suppose C_i is the ith individual classifier in the ensemble system $\{C_1(\boldsymbol{X}), C_2(\boldsymbol{X}), \cdots, C_M(\boldsymbol{X})\}$. $\boldsymbol{X}$ is an input image data set and y is the class label vector corresponding to a given input. C_i is usually obtained by a learning process with the training dataset. A given training dataset $\boldsymbol{X}_{train} \in \Re^{d \times N}$ containing N pixels is usually denoted as $\boldsymbol{X}_{train} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]$. $\boldsymbol{y}_{train} = [y_1, y_2, \cdots, y_N]$ is the true class labels corresponding to $\boldsymbol{X}_{train}$, where $y_i \in \{1, 2, \cdots, K\}$. A classifier can be described as an unknown function $y = f(\boldsymbol{x})$, which can predict the class y for a new input vector $\boldsymbol{x}$. The classifier $C_i(\boldsymbol{x})$ is a hypothesis $f_i(\boldsymbol{x})$ about the true function f. The idea of the ensemble learning system is combining multiple classifiers to make the final decision. The most common way is assigning a weight w_i to each classifier C_i for the weighted majority voting. The ensemble result for $\boldsymbol{x}$ can be denoted as $y = \boldsymbol{w}^T F(\boldsymbol{x})$, where $\boldsymbol{w}$ is the weight vector and $F(\boldsymbol{x})$ is the predicted label vector whose entries generated by the hypotheses $f_i(\boldsymbol{x})$. According to our experience, an ensemble algorithm may reach better performance than a single classifier. However, an ensemble learning system includes two major problems: one is the number of classifier. An excess of classifiers maybe not help for classification, while increase the computational burden in testing stage; the other is how to get a set of appropriate weights, which will direct influence accuracy of classification.

Above two problems can be interpreted as a desired ensemble learning system should use the minimal classifiers while achieve the maximum accuracy, i.e. minimize the difference between the predictions and true labels. It is noted that there are two similar tasks emphasized in sparse recovery process. One is obtain a sparse solution as the weights of a linear combination. The other is minimizing the approximation error between the observations and the recovered sets. Comparatively speaking, the task of selecting the minimal classifiers in ensemble system can be equal to minimizing the number of non-zero elements of the weight vector $\boldsymbol{w}$, i.e. ℓ_0 norm of $\boldsymbol{w}$ if $\boldsymbol{w}$ is the weights of individual classifiers. The second task of the ensemble learning, i.e. maximum accuracy, can be considered as minimizing the approximation error between the predictions and the true labels. In the circumstances, we present a new way to deal with the ensemble learning by using the existing sparse recovery technique.

For each classifier C_i, a predicted label vector $f_i(\boldsymbol{X}_{train})$ can be obtained, where $\boldsymbol{X}_{train} \in \Re^{d \times N}$ is the training dataset containing N samples. Consider M classifiers on the training set $\boldsymbol{X}_{train}$, we can get a matrix $\boldsymbol{F}$:

$$\boldsymbol{F} = \begin{pmatrix} f_1(\boldsymbol{x}_1) f_2(\boldsymbol{x}_1) \cdots f_M(\boldsymbol{x}_1) \\ f_1(\boldsymbol{x}_2) f_2(\boldsymbol{x}_2) \cdots f_M(\boldsymbol{x}_2) \\ \vdots\ \vdots\ \vdots\ \vdots\ \vdots \\ f_1(\boldsymbol{x}_N) f_2(\boldsymbol{x}_N) \cdots f_M(\boldsymbol{x}_N) \end{pmatrix}$$

$\boldsymbol{F} \in \Re^{N \times M}$ is the predicted matrix of the training dataset $\boldsymbol{X}_{train}$. Then we can write the weighted voting as follows:

$$\begin{pmatrix} f_1(\boldsymbol{x}_1) f_2(\boldsymbol{x}_1) \cdots f_M(\boldsymbol{x}_1) \\ f_1(\boldsymbol{x}_2) f_2(\boldsymbol{x}_2) \cdots f_M(\boldsymbol{x}_2) \\ \vdots\ \vdots\ \vdots\ \vdots\ \vdots \\ f_1(\boldsymbol{x}_N) f_2(\boldsymbol{x}_N) \cdots f_M(\boldsymbol{x}_N) \end{pmatrix} \bullet \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \tag{7}$$

In the sparse representation framework, the ensemble problem is formulated as

$$\hat{\boldsymbol{w}} = \arg\min \left|\left|\boldsymbol{w}\right|\right|_0 \quad \text{s.t.} \quad \boldsymbol{F}\boldsymbol{w} = \boldsymbol{y} \quad \text{and} \quad w_i \ge 0 \tag{8}$$

where $y = (y_1, \cdots, y_N)^T$ is the vector of the true ground labels for all samples in the training dataset, $\boldsymbol{w} \in \Re^M$ is the weight vector and its entries must be equal to or greater than zero, and $\|\boldsymbol{w}\|_0$ denotes the number of non-zero entities in vector $\boldsymbol{w}$. The dictionary $\boldsymbol{F}$ is a predicted label matrix whose columns are predicted labels generated by each of the individual classifiers for the training dataset. Let Λ be the index set of nonzero term of $\boldsymbol{w}$, and $\tilde{M}$ is the size of Λ. After obtaining $\boldsymbol{w}$, we can get the predicted label $f_i(\boldsymbol{x}_{test})$ of the testing sample $\boldsymbol{x}_{test} \in \boldsymbol{X}_{test}$ by classifier $C_i(\boldsymbol{x}), i \in \Lambda$. Then, we can get a predicted label vector $(f_1(\boldsymbol{x}_{test}), f_2(\boldsymbol{x}_{test}), \cdots, f_{\tilde{M}}(\boldsymbol{x}_{test}))$ for the testing sample $\boldsymbol{x}_{test}$. The final classification label of the testing sample $\boldsymbol{x}_{test}$ is generated by applying the WMV/MV rules to the predicted labels $(f_1(\boldsymbol{x}_{test}), f_2(\boldsymbol{x}_{test}), \cdots, f_{\tilde{M}}(\boldsymbol{x}_{test}))$.

Furthermore, we observed that there are large homogeneous regions in HSI scene, i.e. the neighboring pixels are consisted of the same types of materials (same class). Previous literatures have proved the positive influence of contextual knowledge on HSI classification [42–44,50–52]. In order to improve the performance of ensemble system, a new sparse ensemble learning method with spatial prior is proposed for classification of HSI, which is an extension and application of the abovementioned general framework. In a new ensemble system, the pixels in a small spatial neighborhood are assumed to share certain same classifiers to minimize the difference between the predictions and the true labels. Through joint sparse representation process, we can obtain the weights of classifiers for ensemble learning. Fig. 1 presents an overview of the proposed method.

Let $\boldsymbol{x}_1 \in \Re^d$ be a pixel, and $\boldsymbol{x}_i$ $(i = 2, \cdots, L)$ be its $L-1$ nearest neighbors in the spatial domain. $\boldsymbol{y}_1 \in \Re^N$ is the vector of the true ground labels for all samples in the training dataset $\boldsymbol{X}_{train} \in \Re^{d \times N}$. Considering M classifiers, learners can get a predicted class label matrix $\boldsymbol{F}_1 \in \Re^{N \times M}$ of the training set $\boldsymbol{X}_{train}$. For the neighbors of the training set, we can get predicted label matrices $\boldsymbol{F}_2, \boldsymbol{F}_3, \cdots, \boldsymbol{F}_L$ in the same way. We assume that $\boldsymbol{y}_i = \boldsymbol{y}_1, i = 2, \cdots, L$, and the underlying sparse classifiers for the neighboring pixels share a common

Please cite this article as: E. Zhang, et al., Spectral–spatial hyperspectral image ensemble classification via joint sparse representation, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.033

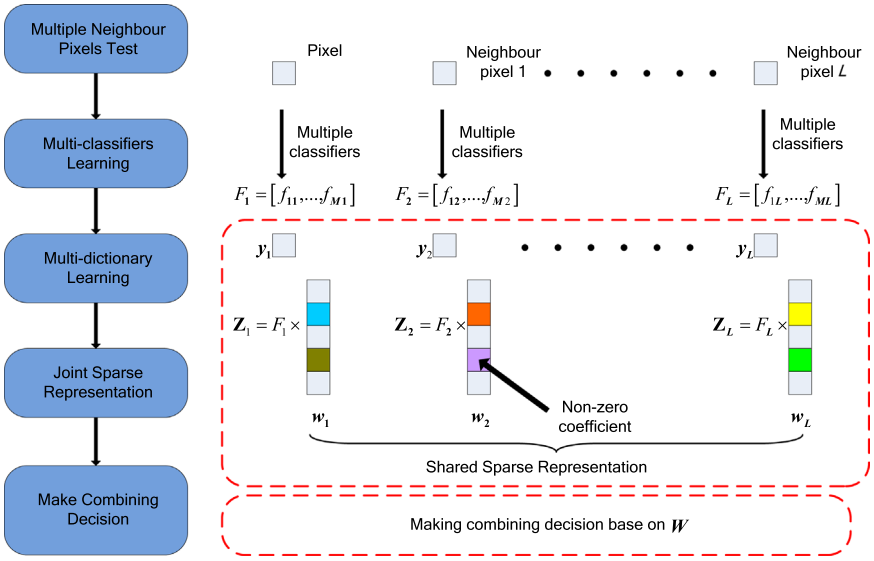

Fig. 1. Overview of the proposed algorithm.

pattern, but these classifiers are weighted with a different set of coefficients. Then, the joint sparse ensemble problem for HSI can be formulated as

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{L} \|\boldsymbol{y}_i - \boldsymbol{F}_i\boldsymbol{w}_i\|_2^2 + \lambda\|\boldsymbol{W}\|_{1,q} \quad \text{s.t.} \quad \boldsymbol{W} \geq 0 \tag{9}$$

where $\boldsymbol{w}_i \in \Re^M$ is the ith column vector of $\boldsymbol{W} \in \Re^{M\times L}$ and its entries must be equal to or greater than zero. λ is a positive parameter and q is set greater than 1 to make the optimization problem convex. Here, $\|\boldsymbol{W}\|_{1,q}$ is a norm defined as $\|\boldsymbol{W}\|_{1,q} = \sum_{k=1}^{M}\|\boldsymbol{W}^k\|_q$ where $\boldsymbol{W}^k$ is the kth row vector of $\boldsymbol{W}$; $\|\boldsymbol{W}^k\|_q = \left(\sum \left|\boldsymbol{W}_i^k\right|^q\right)^{\frac{1}{q}}$. The weights for neighbors are different, thus these weights can be regulated. In our method, we consider the values in matrix W less than $1e^{-4}$ as 0 after getting W. Thus, the weight vector of ensemble learning can be controlled and adjusted to get good ensemble outputs for training samples as well as corresponding neighbor samples.

After obtaining $\boldsymbol{W}$, we can generate a final output for a new testing pixel by combining the decisions of the selected classifiers with their weights. In our experiments, we use w (the first column of $\boldsymbol{W}$) as the weights of joint decision-making, which is denoted as JSWMV. The detailed procedure of the proposed ensemble learning method is given in Algorithm 2.

Algorithm 2. Ensemble system based on joint sparse representation

Input: Given training dataset $\boldsymbol{X}_{train} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N] \in \Re^{d\times N}$, true label of the training dataset $\boldsymbol{y}_{train} = [y_1, y_2, \cdots, y_N]^T$, testing dataset $\boldsymbol{X}_{test} = [\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \cdots, \tilde{\boldsymbol{x}}_{\tilde{N}}] \in \Re^{d\times\tilde{N}}$.

Output: Desired number of classifiers for the ensemble; $\tilde{M}$ classifiers and classification results $\boldsymbol{y}_{test}$ of the testing dataset

Step 1: Construct M individual classifiers $C_i(\boldsymbol{x})$, $i = 1, 2, \cdots, M$ as described Algorithm 1.

Step 2: Generate the predicted label matrices of the training set and corresponding neighbor sets $\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_L$ by the classifiers obtained by Step 1.

Step 3: Apply a compressed sensing method for optimizing the optimal weighting $\boldsymbol{W}$ for the classifiers by formula (9). Here, we propose the alternating direction method of multipliers (ADMM) to solve the optimization problem as Algorithm 3.

Step 4: After getting $\boldsymbol{W}$ from Step 3, vector w is the first column of $\boldsymbol{W}$. Λ is the index set of nonzero terms in w, and $\tilde{M}$ is the size of Λ. We can obtain the predicted labels $f_i(\boldsymbol{x}_{test})$ of the testing sample $\boldsymbol{x}_{test} \in \boldsymbol{X}_{test}$ by classifier $C_i(\boldsymbol{x})$, $i \in \Lambda$. Then, we can get a predicted label vector $(f_1(\boldsymbol{x}_{test}), f_2(\boldsymbol{x}_{test}), \cdots, f_{\tilde{M}}(\boldsymbol{x}_{test}))$ for the testing sample $\boldsymbol{x}_{test}$. In the same way, we can generate a predicted label matrix $(f_1(\boldsymbol{X}_{test}), f_2(\boldsymbol{X}_{test}), \cdots, f_{\tilde{M}}(\boldsymbol{X}_{test}))$ for all data point in the testing dataset.

Step 5: Generate an output for the data point from the testing dataset by combining those decisions of the selected classifiers with their weights w_i, $i \in \Lambda$:

$$\boldsymbol{y}_{test} = (f_1(\boldsymbol{X}_{test}), f_2(\boldsymbol{X}_{test}), \cdots, f_{\tilde{M}}(\boldsymbol{X}_{test}))\bullet(w_1, w_2, \cdots, w_{\tilde{M}})^T$$

3.2. Optimization algorithm

Here, we present an algorithm to solve the formula (9) based on the classical ADMM [53,54]. Let $\Omega(\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^{L}\|\boldsymbol{y}_i - \boldsymbol{F}_i\boldsymbol{w}_i\|_2^2$. In ADMM, the idea is to decouple $\Omega(\boldsymbol{W})$ and $\|\boldsymbol{W}\|_{1,q}$ by introducing auxiliary variables to reformulate the problem into a constrained optimization problem:

$$\min_{\boldsymbol{W},\boldsymbol{V}} \Omega(\boldsymbol{W}) + \lambda\|\boldsymbol{V}\|_{1,q} \quad \text{s.t.} \quad \boldsymbol{W} = \boldsymbol{V} \tag{10}$$

Since formula (10) is an equally constrained problem and the Augmented Lagrangian Method (ALM) can be used to solve the problem, this can be done by minimizing the augmented Lagrangian function $\Gamma_\mu(\boldsymbol{W}, \boldsymbol{V}; \boldsymbol{Z})$ defined as

$$\Omega(\boldsymbol{W}) + \lambda\|\boldsymbol{V}\|_{1,q} + \langle \boldsymbol{Z}, \boldsymbol{W} - \boldsymbol{V}\rangle + \frac{\mu}{2}\|\boldsymbol{W} - \boldsymbol{V}\|_F^2 \tag{11}$$

where $\boldsymbol{Z}$ is the multiplier of the linear constraints. μ is the positive penalty parameter and is updated by $\min(\rho\mu, \max_\mu)$. The ALM algorithm solves $\Gamma_\mu(\boldsymbol{W}, \boldsymbol{V}; \boldsymbol{Z})$ with respect to $\boldsymbol{W}$ and $\boldsymbol{V}$ jointly, keeping $\boldsymbol{Z}$ fixed, and then updating $\boldsymbol{Z}$ keeping the remaining variables fixed. Due to the separable structure of the objective function $\Gamma_\mu(\boldsymbol{W}, \boldsymbol{V}; \boldsymbol{Z})$, one can further simplify the problem by minimizing $\Gamma_\mu(\boldsymbol{W}, \boldsymbol{V}; \boldsymbol{Z})$ with respect to variables $\boldsymbol{W}$and $\boldsymbol{V}$separately. Different steps of the algorithm are given in Algorithm 3, and all the steps are described in more detail below.

Algorithm 3. Alternating direction method of multipliers (ADMM)

Initialization: $\boldsymbol{W}^0, \boldsymbol{V}^0, \boldsymbol{Z}^0, \mu = 10^{-6}$, $\max_\mu = 10^6$, $\rho = 1.1$, maximum iterations $\max_t$.

While stopping criterion has not been met **do**

Please cite this article as: E. Zhang, et al., Spectral–spatial hyperspectral image ensemble classification via joint sparse representation, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.033

Step 1: $\boldsymbol{W}^{t+1}=\text{argmin}_{\boldsymbol{W}}\Gamma_{\mu}(\boldsymbol{W},\boldsymbol{V}^{t};\boldsymbol{Z}^{t})$

Step 2: $\boldsymbol{V}^{t+1}=\text{argmin}_{\boldsymbol{V}}\Gamma_{\mu}(\boldsymbol{W}^{t+1},\boldsymbol{V};\boldsymbol{Z}^{t})$

Step 3: $\boldsymbol{Z}^{t+1}=\boldsymbol{Z}^{t}+\mu(\boldsymbol{W}^{t+1}-\boldsymbol{V}^{t+1})$

Step 4: Check if iteration counter $t>\max_t$ or $\max_{i,j}\left|\boldsymbol{W}_{i,j}^{t+1}-\boldsymbol{V}_{i,j}^{t+1}\right|<10^{-6}$, stop the procedures and output the final sparse coefficients matrix $\boldsymbol{W}$; otherwise, update the parameter μ by $\mu=\min(\rho\mu,\max_{\mu})$.

End while

In Step 1, we need solve a minimization problem of $\Gamma_{\mu}(\boldsymbol{W},\boldsymbol{V};\boldsymbol{Z})$ related to $\boldsymbol{W}$. Due to the quadratic structure, it can be easily solved by setting the first-order derivative equal to zero. The solution presents as follows:

$$\boldsymbol{W}_i^{t+1}=\left(\boldsymbol{F}_i^T\boldsymbol{F}_i+\mu\boldsymbol{I}\right)^{-1}\left(\boldsymbol{F}_i^T\boldsymbol{y}_i+\mu\boldsymbol{V}_i^t-\boldsymbol{Z}_i^t\right) \tag{12}$$

where $\boldsymbol{I}$ is a $M\times M$ identity matrix. $\boldsymbol{W}_i^{t+1}$, $\boldsymbol{V}_i^t$ and $\boldsymbol{Z}_i^t$ are columns of $\boldsymbol{W}^{t+1}$, $\boldsymbol{V}^t$ and $\boldsymbol{Z}^t$, respectively. To make $\boldsymbol{W}_i^{t+1}$equal to or greater than zero, $\boldsymbol{W}_i^{t+1}=\max(\boldsymbol{W}_i^{t+1},0)$.

In Step 2, the optimization problem with respect to $\boldsymbol{V}$ can be formulated as

$$\min_{\boldsymbol{V}}\frac{1}{2}\left\|\boldsymbol{W}^{t+1}+\frac{1}{\mu}\boldsymbol{Z}^t-\boldsymbol{V}\right\|_F^2+\frac{\lambda}{\mu}\|\boldsymbol{V}\|_{1,q} \tag{13}$$

Due to the separable structure of formula (13), we can solve it by minimizing with respect to each row of $\boldsymbol{V}$ separately. Let $\boldsymbol{\omega}_i^{t+1}$, $\boldsymbol{z}_i^{t+1}$ and $\boldsymbol{v}_i^{t+1}$ be rows of matrices $\boldsymbol{W}^{t+1}$, $\boldsymbol{Z}^{t+1}$ and $\boldsymbol{V}^{t+1}$ respectively. For $i=1,\cdots,M$, we can solve the following sub-problem:

$$\boldsymbol{v}_i^{t+1}=\text{argmin}_{\boldsymbol{v}}\frac{1}{2}\left\|\boldsymbol{\omega}_i^{t+1}+\frac{1}{\mu}\boldsymbol{z}_i^{t+1}-\boldsymbol{v}\right\|_2^2+\frac{\lambda}{\mu}\|\boldsymbol{v}\|_q \tag{14}$$

Let $\Delta=\boldsymbol{\omega}_i^{t+1}+\frac{1}{\mu}\boldsymbol{z}_i^{t+1}$ and $\eta=\frac{\lambda}{\mu}$. One can deduce the solution of formula (14) for any q. In this paper, we only focus on the case where $q=2$. The solution of formula (14) follows as

$$\boldsymbol{v}_i^{t+1}=\Psi\left(1-\frac{\eta}{\|\Delta\|_2}\right)\cdot\Delta \tag{15}$$

where $\Psi(\delta)=\max(\delta,0)$.

The adopted ADMM optimization method is computationally efficient to solve the joint sparsity constraint problem. The main steps of the algorithm are the update steps of $\boldsymbol{W}$and $\boldsymbol{V}$. The update step for $\boldsymbol{W}$ involves computing $\left(\boldsymbol{F}_i^T\boldsymbol{F}_i+\mu\boldsymbol{I}\right)^{-1}$ and three matrix multiplications. The $\left(\boldsymbol{F}_i^T\boldsymbol{F}_i+\mu\boldsymbol{I}\right)^{-1}$ is constant across iterations and can be pre-computed. Matrix multiplication for two matrices of sizes $M\times N$ and $N\times 1$ can be done in $O(MN)$. Hence, for a given training and test data, the computations are linear in the size of test data. Update step for $\boldsymbol{V}$ involves only scalar matrix computations and is very fast. Classification step is similar as the traditional weighted voting, and is efficient.

4. Experiments and performance comparisons

In this section, we will evaluate the proposed joint sparse representation-based ensemble learning classification method on three hyperspectral images.

4.1. Data sets

The first one is Indian Pines image obtained by the Airborne Visible/Infrared Imaging Spectrometer. It includes a wide range of wavelengths from 0.2 to 2.4 μm and has spatial resolution of 20 m. In our experiments, we cut a part scene of 145×145, and choose the commonly-used 200 bands for testing. This scene contains 16 classes (Fig. 2(a)), most of which are different types of crops (e.g., corns, soybeans, and wheat).

The second test image is the University of Pavia acquired by the Reflective Optics System Imaging Spectrometer (ROSIS), which is an urban image with 1.3 m spatial resolution and a spectral range from 0.43 to 0.86 μm. The size of this image is $610\times 340\times 103$ after removing 12 noisy bands. The ground-truth image contains 9 classes, as shown in Fig. 2(b).

In our experiments, the other scene gathered by the ROSIS sensor is the Center of Pavia. This scene consists of 1096×492 pixels and 102 spectral bands after removing 13 noisy bands. The reference image of 9 classes is shown in Fig. 2(c).

4.2. Comparison of classification results

In our experiments, the classification and regression tree (CART) [45,46] is used as the base classifier. CART is only one

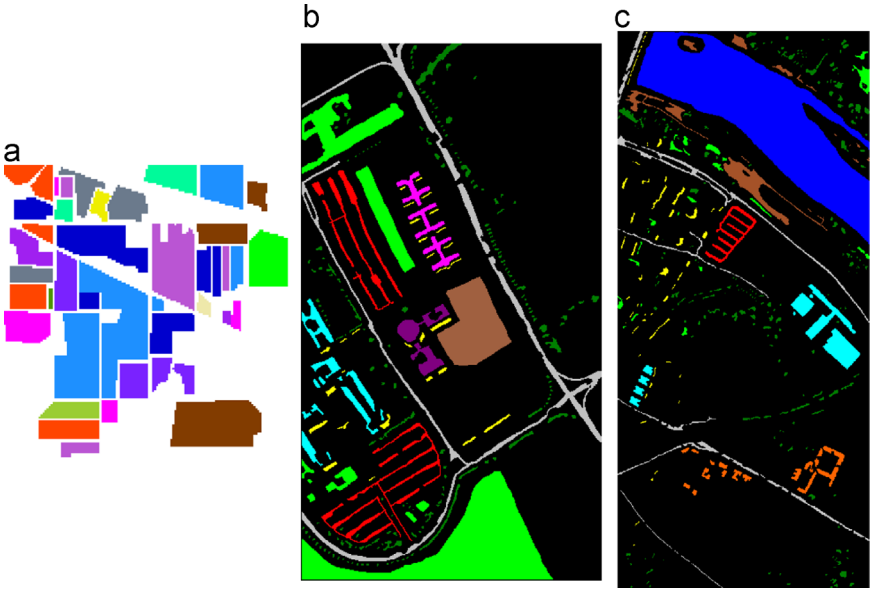

Fig. 2. Reference image: (a) Indian Pines image; (b) University of Pavia image; and (c) Center of Pavia image.

Please cite this article as: E. Zhang, et al., Spectral–spatial hyperspectral image ensemble classification via joint sparse representation, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.033

example of a more generic tree-growing methodology and provides a general framework that can be instantiated in various ways to produce different decision trees. Thus, it is always used in ensemble system. For each image, the results of the classical CART with original features are used as baseline in our experiments denoted as "CART". Random subspace strategy is used to produce a set of different individual classifiers. Specifically, we use randomly selected band subsets of a training dataset to train each individual CART classifier. In experiments, ensemble system with fusion decision criterion (1), (3) and (4) is represented as MV, WMV1, and WMV2 respectively. The ensemble system with fusion decision criterion (8) is denoted as SWMV which does not consider neighborhood information. The ensemble system using joint sparse representation coefficients as weights is denoted as JSWMV. The classical CART and the ensembles with different combination rules are compared qualitatively and quantitatively on the test set. In quantitative aspect, the overall accuracy (OA), average accuracy (AA), the kappa coefficient measure (Kappa), and the associated standard deviation are adopted to evaluate the effectiveness of different classifiers. All the programs are executed using MATLAB 2010 in the environment of an Intel Core i3-550 CPU 3.20 GHz and 4 GB of RAM. In each case, we show the running time of the test stage for comparing computational efficiency.

We can get neighbor samples through various methods ($l \times l$ window or K-nearest neighbor, etc.). For convenience, we use fixed 4-connected neighborhood ($L = 5$) for all datasets shown in Fig. 3. In the classifier ensembles, we use different data subsets to train each individual classifier. 10% to 90% bands of the original bands are selected randomly to form data subsets. Each experiment is performed with three different ensemble sizes, 100, 300 and 500 classifiers. And, we do the experiment 50 times to get the average results. The classification accuracies in bold represent the best results of corresponding methods.

The first experiment is performed on the Indian Pines image. For each of the 16 classes, about 5% samples are selected for training and the remaining samples for testing. It is a challenging classification task for the traditional single classifier because few training samples can be used to learn [55]. The results of ensemble system with 500 individual classifiers are shown in Table 1. The classification maps generated by various techniques are shown in Fig. 4. The result of the classical CART with original features for training is shown in Fig. 4(a). The maps in Fig. 4(b)–(f) show the results of various ensemble learning systems with different combination rules. Fig. 5 evaluates the performances of classifiers produced by 50 times experiments. Box plot is used to show the statistical results, which represents the lower quartile, median and upper quartile values in lines. The X-axis represents the name of classification methods. The Y-axis is the overall accuracies of classifiers.

First of all, it can be seen that the classification results of all ensemble learning systems are much better than those of single classifier. It demonstrates that the combination of multiple classification results is a good way to solve some challenging classification tasks. From Fig. 4, the result maps of MV (b), WMW1 (c), and WMV2 (d) methods are very similar. However, through careful observation (for example, regions labeled by the black lined box in Fig. 4), we can see that the results of JSWMV methods are better than those of MV, WMV1, WMV2 and SWMV. From Table 1, it can be seen that the performance of traditional ensemble learning is better than the one of single classifier. The corresponding gain in accuracy with respect to single CART is about 8%. However, JSWMV shows a much better performance. The average accuracy improvement compared to CART is about 11%. It is noted that the 9th class (Oats) has 20 total samples, where only one sample for training. It is a challenging task for most classification method. In this situation, traditional ensemble and the proposed JSWMV have worse performance than a single classifier. Because single classifier may give a right decision with a low frequency, while the ensemble including too many "bad" individuals would make a wrong decision with a great probability. JSWMV will inevitably encounter the similar problems. Besides, JSWMV may make light of the reconstruction error of one element in a high-dimensional signal, which leads to poor performance on the 9th class. In order to further discuss this problem, we increase the number of training data from 1 to 5 for the 9th class. Although the number of 9th class increases, it still exists the imbalanced problem. The experimental results are shown in Table 2. As shown in Table 2, single classifier outperforms the ensemble methods when there are 1 or 2 training samples. But the accuracy of JSWMV is no longer 0. When the number of training samples increases to 3, the accuracy of the single classifier has a big improvement. Moreover, all the ensemble methods outperform the single classifier. As the number of training samples increases to 4, JSWMV outperforms the other compared approaches in terms of accuracy.

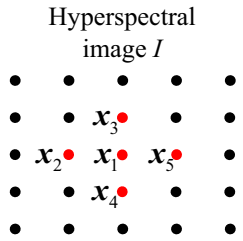

Fig. 3. Four nearest neighbors of a pixel x_1.

Secondly, SWMV and JSWMV are sparse ensemble methods. The weight values of SWMV and JSWMV can be obtained from formulas (8) and (9), respectively. Comparing results of SWMV and JSWMV, the corresponding gain in accuracy is about 2%. It proves that the proposed methods considering neighborhood knowledge can select a preferable subset of classifiers in ensemble learning. It also demonstrates that joint sparse representation coefficients can be used as weights and have a better performance to those of classical combination rules. Moreover, JSWMV can complete the selection of classifiers subset and obtain weights in one step.

Thirdly, the overall accuracies of 50 runs for various methods are tabulated in Fig. 5. It can be observed that the proposed JSWMV method outperforms the other compared approaches in terms of OA and is with good consistency at all times. Thus, the proposed JSWMV are promising ensemble method for HSI classification.

The second and third experiments are conducted on the University of Pavia and Center of Pavia images, respectively. For the University of Pavia image, 1% samples selected randomly from each class are used in training and the remaining 99% of data is used for testing. To make the classification for the Center of Pavia image more challenging, only 15 labeled samples from each class are randomly selected as training samples and the remainder as the test samples. The quantitative results of the ensemble system with 500 individual classifiers are shown in Tables 3 and 4. The classification maps obtained by various classifiers are shown in Figs. 6 and 8. Corresponding box plots evaluating the consistency of 50 times experiments are shown in Figs. 7 and 9, respectively. As can be observed from Tables 3 and 4, similar to the results on the Indian Pines image, ensemble methods lead to much improvement compared with the single classifier methods. In these two images, the proposed JSWMV performs better than the single classifier method, traditional ensemble methods (MV, WMV1 and WMV2) as well as SWMV in terms of visual quality of classification maps and quantitative metrics (OA, AA and Kappa). Note that, the gains of the JSWMV are more than 2–4%. Overall, those experiments also prove that the proposed JSWMV can

Please cite this article as: E. Zhang, et al., Spectral–spatial hyperspectral image ensemble classification via joint sparse representation, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.033

Table 1
Classification accuracy (%) for the Indian Pines image on the test set.

Class	# Samples		CART	MV	WMV1	WMV2	SWMV	JSWMV
	Training	Testing						
1	3	51	29.92	37.92	37.57	37.49	40.24	**42.00**
2	72	1362	52.02	61.54	61.16	61.04	61.20	**63.38**
3	42	792	38.70	48.20	48.03	47.91	47.75	**49.59**
4	12	222	34.62	42.81	42.72	42.65	41.82	**43.35**
5	25	472	71.65	75.43	75.28	75.20	76.32	**76.86**
6	38	709	83.56	89.13	89.05	88.92	89.04	**90.49**
7	2	24	12.25	15.17	15.17	15.50	**16.92**	14.50
8	25	464	86.98	91.69	91.70	91.56	92.23	**93.13**
9	1	19	**2.84**	1.05	1.05	1.05	0.21	0
10	49	919	54.73	63.70	63.52	63.42	64.03	**64.46**
11	123	2345	64.01	75.85	75.98	75.89	76.01	**77.77**
12	31	583	36.07	43.55	43.78	43.76	45.73	**46.55**
13	11	201	80.55	92.81	92.63	92.51	93.51	**94.75**
14	65	1229	88.46	91.99	91.88	91.72	92.88	**94.02**
15	19	361	42.47	45.47	45.74	45.70	45.96	**46.74**
16	5	90	77.58	82.16	82.16	82.11	84.53	**85.82**
OA			61.74 ± 1.34	70.18 ± 0.98	70.14 ± 0.98	70.02 ± 0.97	70.77 ± 1.34	$\mathbf{72.82 \pm 0.81}$
AA			53.03 ± 2.67	59.06 ± 2.49	59.03 ± 2.49	58.93 ± 2.48	58.58 ± 2.32	$\mathbf{61.40 \pm 2.13}$
Kappa			56.45 ± 1.56	65.87 ± 1.13	65.82 ± 1.13	65.69 ± 1.13	65.21 ± 1.50	$\mathbf{68.86 \pm 0.95}$
Num			1	500	500	500	44.33	72.82
Time (s)			0.64	7.17	7.46	7.55	1.84	11.31

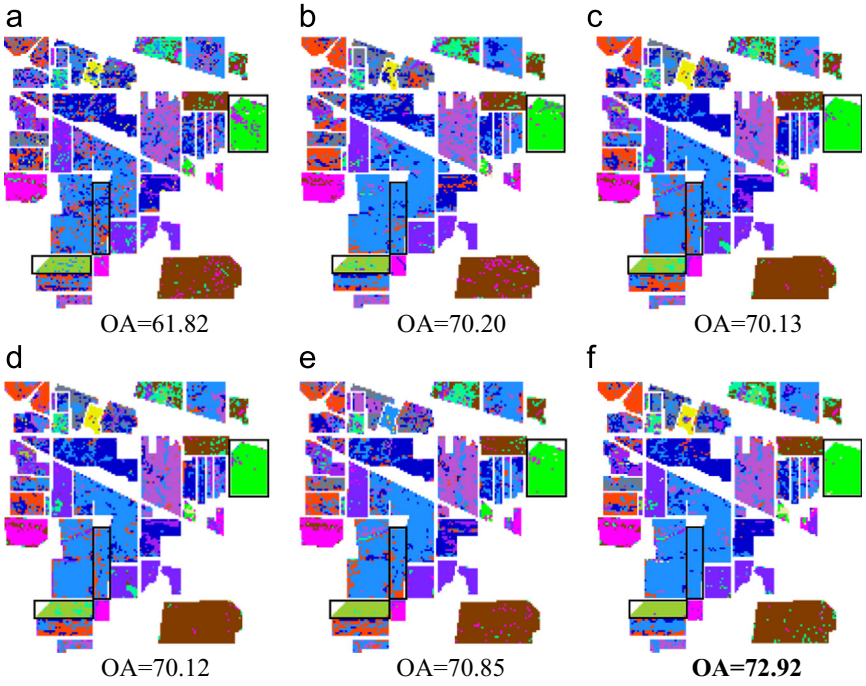

Fig. 4. For the Indian Pines images: classification maps obtained by (a) CART, (b) MV, (c) WMV1, (d) WMV2, (e) SWMV, and (f) JSWMV.

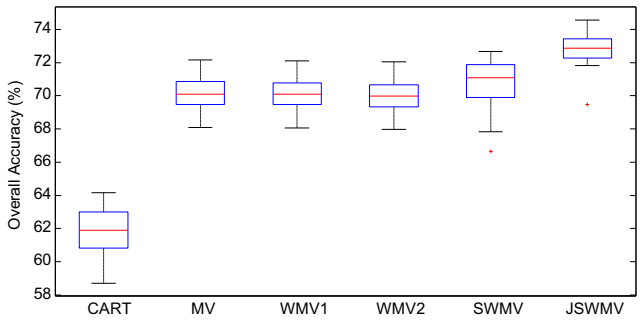

Fig. 5. Box plot of the classification accuracies for the Indian Pines image.

achieve a good classification performance with a much smaller number of classifiers.

In addition, experiments not only indicate that it is necessary to do the classifier selection, but also show that the proposed method is effective in the aspect of selecting classifiers. The numbers of classifiers (Num) in various methods are shown in Tables 1, 3 and 4. Although the SWMV algorithm selects classifiers less than JSWMV, the proposed JSWMV not only reduces the classifiers (less than 500) but also preserves useful information of HSI to provide a better classification results. In the case of the Indian Pines image, the number of nonzero entries in $\boldsymbol{w}$ obtained by JSWMV ranges from 60 to 80. We randomly obtain a set of weight coefficients $\boldsymbol{w}$ in which there are 65 nonzero values as shown in Fig. 10. That

Please cite this article as: E. Zhang, et al., Spectral–spatial hyperspectral image ensemble classification via joint sparse representation, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.033

means that only 65 classifiers are selected to make the final decision in this ensemble system. The proposed JSWMV method gets much smaller number of nonzero weights than 500, which means that JSWMV uses a small part of classifiers to predict the class labels of the testing samples and make the final decision. Thus, JSWMV dramatically reduces test time when there are large ensembles or large number of unlabeled pixels to be classified.

For the running time comparison, the detailed average running times for each case of each classifier is shown in Tables 1, 3 and 4. SWMV and JSWMV with the selection operation of classifiers need some time to compute the ensemble weights. But, SWMV use comparative or less time than the traditional ensemble method. The reason is that SWMV chooses a few classifiers in the final ensemble, which reduces the computational time for dealing with the unlabeled test examples. In contrast with SWMV, JSWMV takes advantage of the extended neighborhood information, which can improve the accuracy but also increase the computational load simultaneously. But the number of neighborhoods usually is small, which makes a relatively small influence on the computational load. Furthermore, we can use some existing efficient sparse optimization method [30], e.g. Accelerated Proximal Gradient [31], to solve this problem.

Table 2
Classification accuracy (%) for the 9th class in Indian Pines image on the test set.

Num. of class 9	CART	MV	WMV1	WMV2	SWMV	JSWMV
1	**2.84**	1.05	1.05	1.05	0.21	0
2	**16.00**	12.11	12.22	12.44	9.00	9.67
3	38.82	43.65	43.76	43.76	**43.88**	42.94
4	43.75	58.25	58.25	57.87	59.62	**64.00**
5	44.93	58.67	58.67	58.13	64.53	**69.87**

4.3. Effects of parameters

Here, we further analyze the effects of parameters on the three images.

Firstly, we discuss the effects of the number of neighborhood to the proposed method. The number of neighborhoods increases from 4 to 8, i.e., all the neighbor samples in the 3×3 window are used. We test JSWMV with fixed 8-connected neighborhood on the Indian Pines, University of Pavia, and Center of Pavia images. The experimental setup is the same as the above experiments. The quantitative results of the ensemble system are shown in Tables 5, 6 and 7 respectively. Experimental results on three hyperspectral images show that neighborhood information is helpful for classification. JSWMV outperforms SWMV which does not consider neighborhood information. However, large neighborhood may

Table 3
Classification accuracy (%) for the University of Pavia image on the test set.

Class	# Samples		CART	MV	WMV1	WMV2	SWMV	JSWMV
	Training	Testing						
1	66	6565	74.29	82.54	82.48	82.33	83.28	**85.98**
2	186	18,463	84.83	91.98	91.93	91.82	92.74	**93.89**
3	20	2079	49.25	52.93	52.92	52.99	53.24	**54.66**
4	30	3034	73.04	77.55	77.57	77.46	78.78	**79.55**
5	13	1332	93.09	95.66	95.65	95.54	96.65	**96.87**
6	50	4979	43.88	43.83	43.92	43.94	43.98	**45.56**
7	13	1317	43.65	49.24	49.30	49.28	52.56	**55.08**
8	37	3645	54.17	65.01	65.01	64.82	68.87	**72.28**
9	10	937	78.66	86.28	86.22	85.91	87.81	**91.16**
OA			71.99 ± 1.62	78.25 ± 1.27	78.23 ± 1.27	78.13 ± 1.26	79.32 ± 1.19	$\mathbf{80.99 \pm 0.98}$
AA			66.10 ± 2.59	71.67 ± 2.32	71.67 ± 2.31	71.57 ± 2.31	73.10 ± 1.55	$\mathbf{75.00 \pm 1.61}$
Kappa			62.69 ± 2.01	70.59 ± 1.73	70.57 ± 1.73	70.44 ± 1.72	72.03 ± 1.58	$\mathbf{74.24 \pm 1.35}$
Num			1	500	500	500	36.48	75.82
Time (s)			1.76	16.06	16.87	17.22	9.39	15.62

Table 4
Classification accuracy (%) for the Center of Pavia image on the test set.

Class	# Samples		CART	MV	WMV1	WMV2	SWMV	JSWMV
	Training	Testing						
1	15	65,263	93.29	95.16	95.09	94.68	95.21	**97.78**
2	15	6493	68.97	74.19	73.94	73.43	73.14	**78.21**
3	15	2890	74.21	76.54	76.51	76.39	75.97	**80.04**
4	15	2125	59.17	64.16	64.06	63.63	62.28	**73.45**
5	15	6534	74.97	**77.35**	77.32	76.94	75.75	76.96
6	15	7570	69.21	76.79	76.70	75.58	73.57	**84.45**
7	15	7272	74.42	76.65	76.62	76.16	75.83	**80.91**
8	15	3107	91.27	93.06	93.06	92.78	92.99	**95.80**
9	15	2150	80.67	85.72	85.69	85.28	86.01	**93.22**
OA			85.96 ± 3.00	88.65 ± 2.20	88.58 ± 2.25	88.12 ± 2.50	88.17 ± 2.36	$\mathbf{91.92 \pm 1.09}$
AA			76.24 ± 2.28	79.96 ± 1.96	79.89 ± 2.02	79.43 ± 2.10	78.97 ± 2.30	$\mathbf{84.53 \pm 1.37}$
Kappa			76.88 ± 4.34	81.07 ± 3.32	80.96 ± 3.38	80.25 ± 3.70	80.28 ± 3.57	$\mathbf{86.31 \pm 1.76}$
Num			1	500	500	500	26.73	34.48
Time (s)			0.19	33.09	34.94	34.98	28.49	35.42

Please cite this article as: E. Zhang, et al., Spectral–spatial hyperspectral image ensemble classification via joint sparse representation, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.033

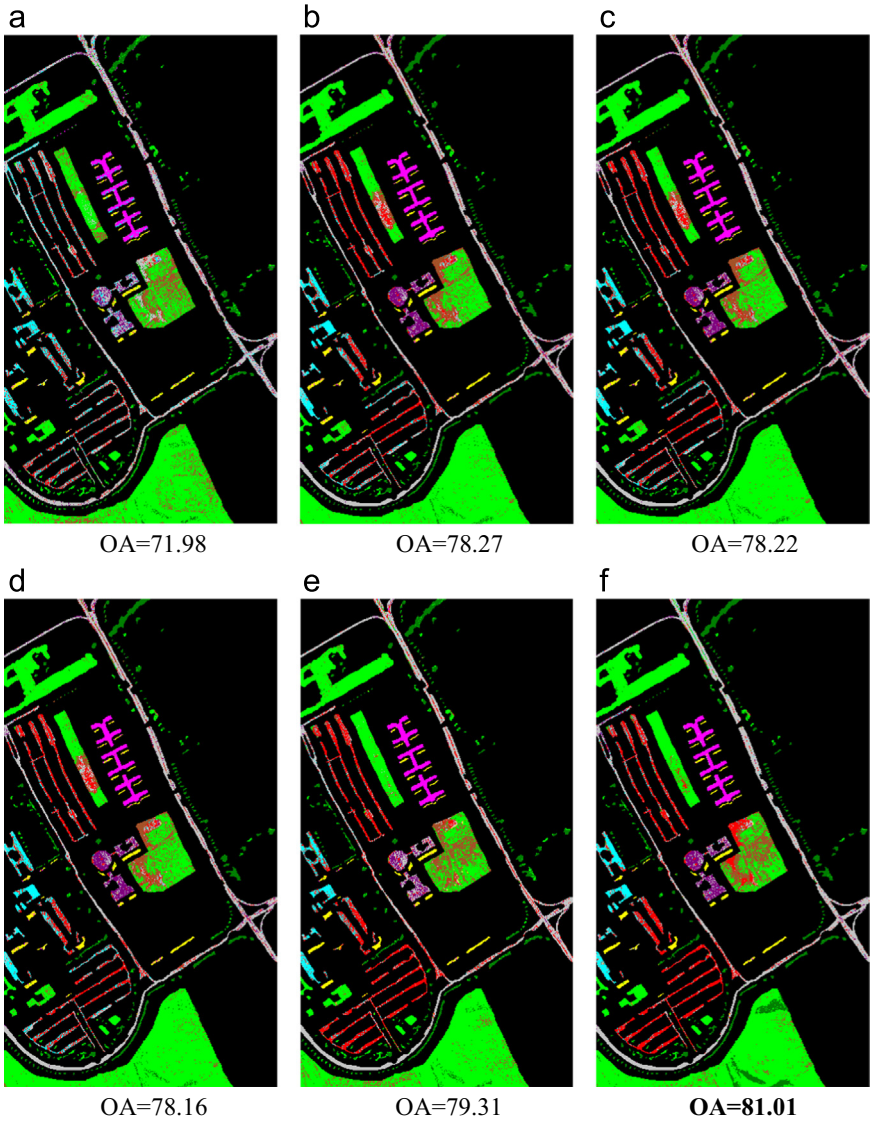

Fig. 6. For the University of Pavia images: (a) CART; (b) MV; (c) WMV1; (d) WMV2; (e) SWMV; and (f) JSWMV.

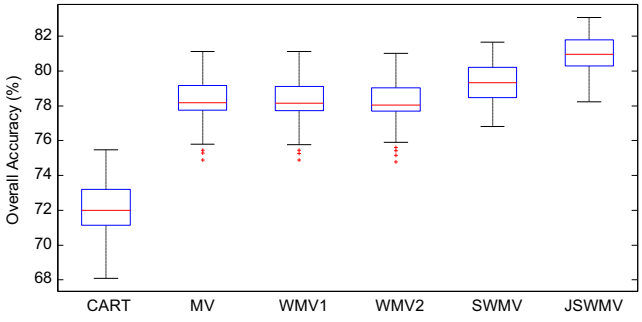

Fig. 7. Box plot of the classification accuracies for the University of Pavia image.

cause oversmoothing over neighboring classes, especially for the pixels near boundaries, which would lead to a decrease in the overall classification accuracy. Along with the increase in the number of neighborhoods from 4 to 8, the changes in classification accuracy are not obvious in our experiments. This indicated that the proposed method has good robustness within a proper neighbor range.

Secondly, we discuss the effect of the number of classifiers to the MV, WMV1, WMV2, SWMV and the proposed JSWMV. In experiments, we randomly select a set of training samples and its corresponding testing sample set. We test all ensemble methods with different numbers of classifiers on the Indian Pines, University of Pavia, and Center of Pavia images. Each experiment is performed with different ensemble sizes ranging from 100 to 1000. The overall accuracies of different classification approaches with different numbers of classifiers are shown in Figs. 11–13. The reported accuracy values are the averaged results over 50 runs. As can be observed in the figures, along with the increase in the number of classifiers from 300 to 1000, the changes in classification accuracy of JSWMV are not obvious, somewhere around 0.5% in our experiments. The experimental results verify the conclusion that an excess of classifiers maybe not help for classification. In addition, the proposed JSWMV method can consistently outperform the other approaches on all the cases.

Finally, we discuss the effect of joint sparsity constraint in formula (9) on the classification performance. The parameter $\lambda > 0$ is used to balance the effects of the two parts in problem (9). In general, the choice of this parameter depends on the prior

Please cite this article as: E. Zhang, et al., Spectral–spatial hyperspectral image ensemble classification via joint sparse representation, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.033

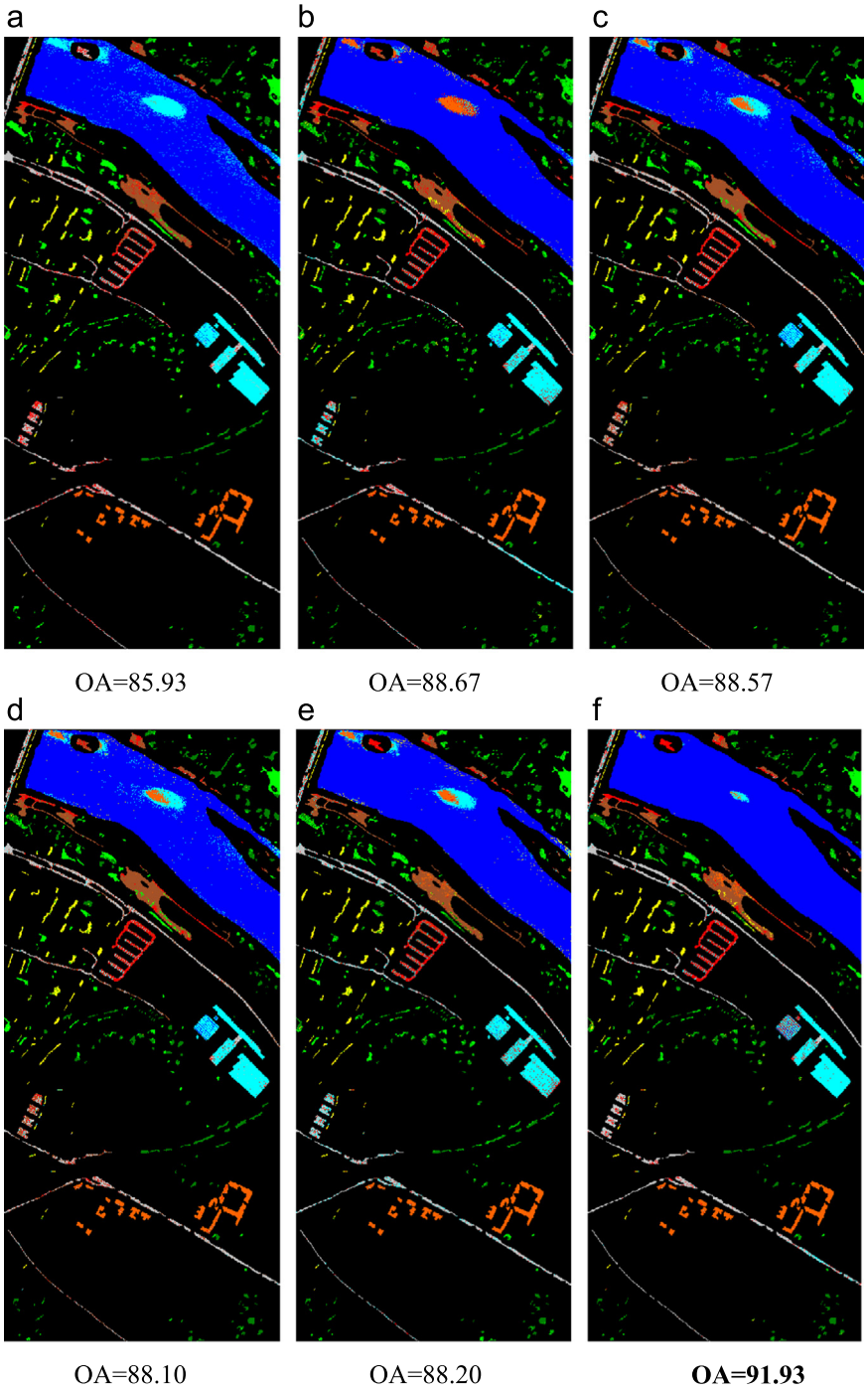

Fig. 8. For the Center of Pavia images: (a) CART; (b) MV; (c) WMV1; (d) WMV2; (e) SWMV; and (f) JSWMV.

knowledge. In experiments, the proposed JSWMV algorithm runs with different values of λ. Fig. 14 shows the overall accuracy variation of classification results across λ values for the three images. All the curves show a sharp increase in performance with λ changing from 0 to 0.005. The classification accuracy is slightly improved, but changes in classification accuracy are not obvious since 0.005. This indicates that imposing joint sparsity constraint is important for the selection of classifier. Moreover, it helps in regulating ensemble performance, when the reconstruction error alone is not sufficient to distinguish between different classifiers. In our experiments, the value of λ is set to 0.01.

4.4. Convergence property of ADMM

The convergence property of ADMM optimization method has been proved in references [53,54]. In Algorithm 3, if iteration counter $t > \max_t$ or $\max_{i,j}\left|\boldsymbol{W}_{i,j}^{t+1} - \boldsymbol{V}_{i,j}^{t+1}\right| < 10^{-6}$, stop the procedures and output the final sparse coefficients matrix $\boldsymbol{W}$. We prove the convergence of our method by experiments. From Fig. 15, it can be seen that the value of $\max_{i,j}|\boldsymbol{W}_{i,j} - \boldsymbol{V}_{i,j}|$ decreases as iteration progress. Fig. 15 intuitively proves the convergence of the optimization problem we need to solve.

Please cite this article as: E. Zhang, et al., Spectral–spatial hyperspectral image ensemble classification via joint sparse representation, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.033

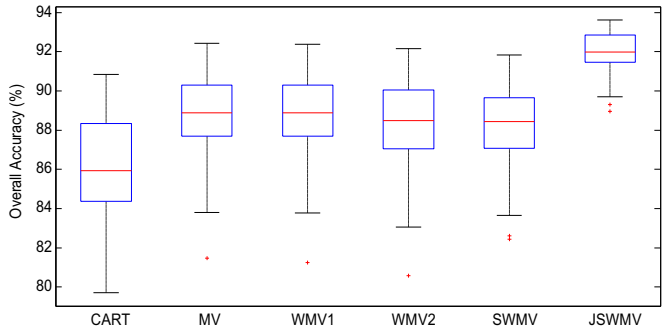

Fig. 9. Box plot of the classification accuracies for the Center of Pavia image.

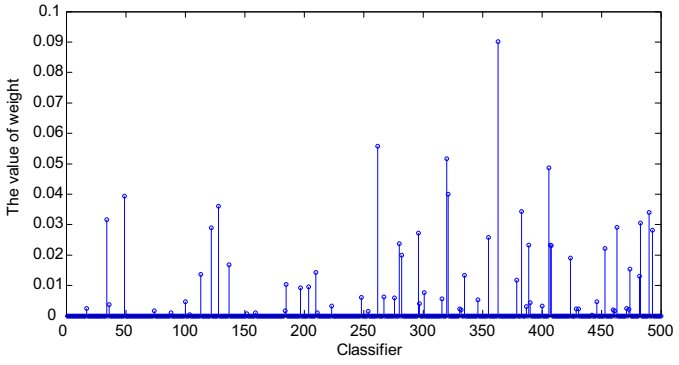

Fig. 10. The value of weight $\boldsymbol{w}$.

Table 5
Classification accuracy (%) for the Indian Pines image on the test set.

	SWMV	JSWMV (4)	JSWMV (8)
OA	70.77 ± 1.34	72.82 ± 0.81	**72.87 ± 0.80**
AA	58.58 ± 2.32	61.40 ± 2.13	**61.78 ± 1.69**
Kappa	65.21 ± 1.50	68.86 ± 0.95	**69.92 ± 0.91**
Num	44.33	72.82	47

Table 6
Classification accuracy (%) for the University of Pavia image on the test set.

	SWMV	JSWMV (4)	JSWMV (8)
OA	79.32 ± 1.19	**80.99 ± 0.98**	80.51 ± 1.02
AA	73.10 ± 1.55	**75.00 ± 1.61**	74.94 ± 1.66
Kappa	72.03 ± 1.58	**74.24 ± 1.35**	73.68 ± 1.39
Num	36.48	75.82	16.70

Table 7
Classification accuracy (%) for the Center of Pavia image on the test set.

	SWMV	JSWMV (4)	JSWMV (8)
OA	88.17 ± 2.36	**91.92 ± 1.09**	91.78 ± 1.17
AA	78.97 ± 2.30	84.53 ± 1.37	**84.67 ± 1.24**
Kappa	80.28 ± 3.57	**86.31 ± 1.76**	86.10 ± 1.88
Num	26.73	34.48	33.94

5. Conclusion

Inspired by the idea of sparse ensembles, we propose a new ensemble learning algorithm based on joint sparse representation for HSI classification. The new approach deals with ensemble learning problems by sparse reconstruction methods, which

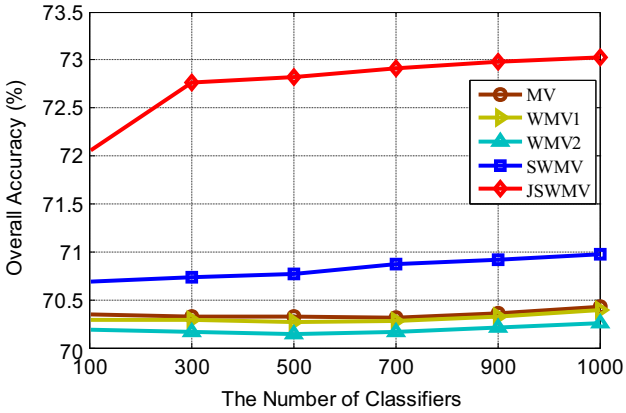

Fig. 11. Effect of the number of classifiers in ensemble learning system for Indian Pines image.

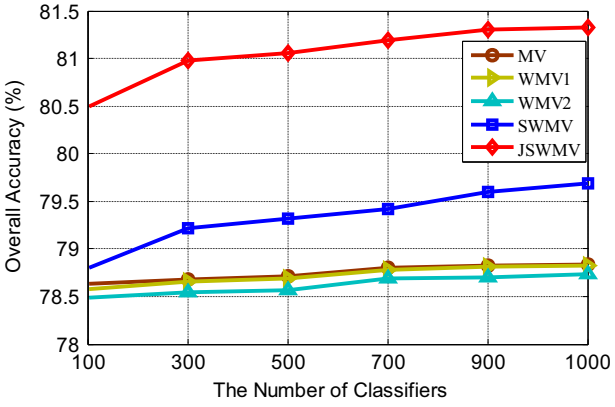

Fig. 12. Effect of the number of classifiers in ensemble learning system for University of Pavia image.

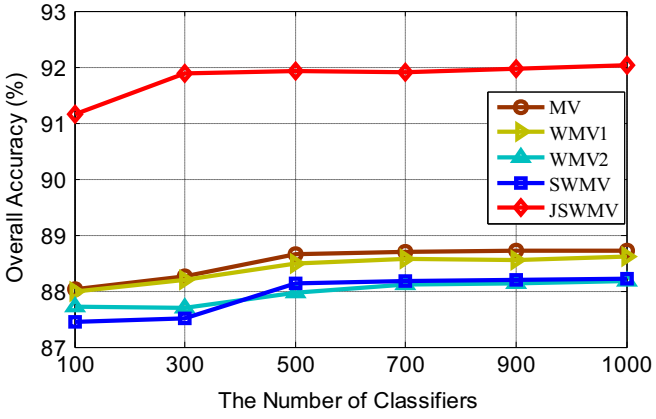

Fig. 13. Effect of the number of classifiers in ensemble learning system for Center of Pavia image.

combines the outputs of individual classifiers with a sparse weight vector. To improve the classification performance, we incorporate the contextual neighborhood knowledge into the sparse representation framework with joint sparse constraint. A set of sparse representation coefficients can be obtained through a sparse recovery process of labels of training pixels and their neighboring pixels. Then, we can use the nonzero term of the coefficients to select classifiers of ensemble learning system, or use the coefficients as weights directly in making a joint decision. Our algorithm can be

Please cite this article as: E. Zhang, et al., Spectral–spatial hyperspectral image ensemble classification via joint sparse representation, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.033

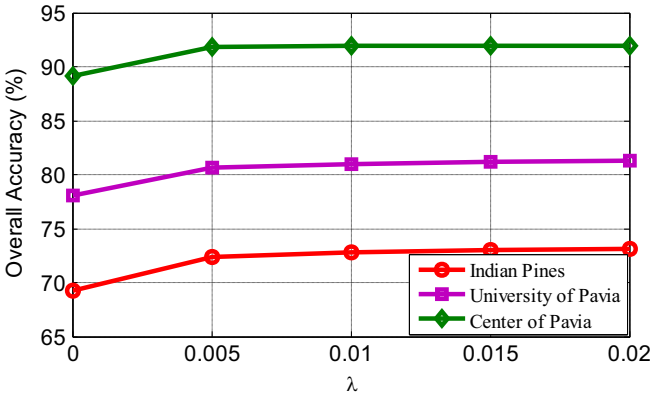

Fig. 14. Effect of the weighting factor λ of the Laplacian constraint for Indian Pines, University of Pavia, and Center of Pavia images.

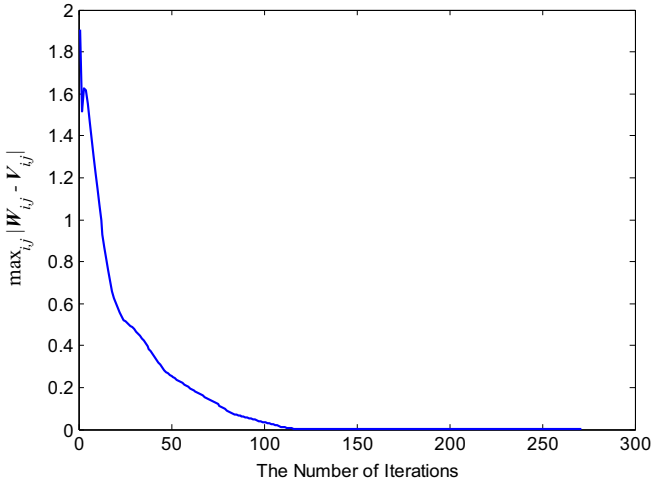

Fig. 15. The convergence property of ADMM optimization method.

applied to the ensemble of any classifiers. In experiments, we adopt CART as the individual classifier and make a detailed comparison of multiple types of combination rule. Experimental results on three hyperspectral images show that not all classifiers contribute in the ensemble system, and the proposed algorithm not only yields better performance than traditional ensemble learning methods but also reduces computational burden in testing phase because of using fewer classifiers. In addition, this paper gives a new way to get effective weights for weighted majority voting.

Conflict of interest

None declared.

Acknowledgment

This work was supported by the National Basic Research Program of China (973 Program, No. 2013CB329402), the National Natural Science Foundation of China (Nos. 61272282, 61203303, 61272279, 61373111, 31300473, and 61501353), the Program for New Century Excellent Talents in University (NCET-13-0948), and the Program for New Scientific and Technological Star of Shaanxi Province (No. 2014KJXX-45).

References

[1] M.T. Eismann, A.D. Stocker, N.M. Nasrabadi, Automated hyperspectral cueing for civilian search and rescue, Proceed. IEEE 97 (2009) 1031–1055.
[2] D. Manolakis, G. Shaw, Detection algorithms for hyperspectral imaging applications, IEEE Signal Process. Mag. 19 (2002) 29–43.
[3] B. Du, Y.X. Zhang, L.P. Zhang, L.F. Zhang, A hypothesis independent subpixel target detector for hyperspectral images, Signal Process. 110 (2015) 244–249.
[4] N. Patel, C. Patnaik, S. Dutta, A. Shekh, A. Dave, Study of crop growth parameters using airborne imaging spectrometer data, Int. J. Remote Sens. 22 (2001) 2401–2411.
[5] B. Datt, T.R. McVicar, T.G. Van Niel, D.L.B. Jupp, J.S. Pearlman, Preprocessing EO-1 Hyperion hyperspectral data to support the application of agricultural indexes, IEEE Trans. Geosci. Remote Sens. 41 (2003) 1246–1259.
[6] B. Hörig, F. Kühn, F. Oschütz, F. Lehmann, HyMap hyperspectral remote sensing to detect hydrocarbons, Int. J. Remote Sens. 22 (2001) 1413–1422.
[7] G. Giacinto, F. Roli, Ensembles of neural networks for soft classification of remote sensing images, in: Proceedings of the European Symposium on Intelligent Techniques, 1997, pp. 20–21.
[8] G.J. Briem, J.A. Benediktsson, J.R. Sveinsson, Multiple classifiers applied to multisource remote sensing data, IEEE Trans. Geosci. Remote Sens. 40 (2002) 2291–2299.
[9] X.F. Song, L.C. Jiao, S.Y. Yang, X.R. Zhang, F.H. Shang, Sparse coding and classifier ensemble based multi-instance learning for image categorization, Signal Process. 93 (2013) 1–11.
[10] J.A. Benediktsson, P.H. Swain, Consensus theoretic classification methods, IEEE Trans. Syst. Man Cybern. 22 (1992) 688–704.
[11] A.B. Santos, A. de Albuquerque Araujo, D. Menotti, Combining multiple classification methods for hyperspectral data interpretation, IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 6 (2013) 1450–1459.
[12] D.S. Frossyniotis, A. Stafylopatis, A multi-SVM classification system, Mult. Classif. Syst. 2096 (2001) 198–207.
[13] X. Ceamanos, B. Waske, J.A. Benediktsson, J. Chanussot, M. Fauvel, J.R. Sveinsson, A classifier ensemble based on fusion of support vector machines for classifying hyperspectral data, Int. J. Image Data Fus. 1 (2010) 293–307.
[14] X. Huang, L. Zhang, An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery, IEEE Trans. Geosci. Remote Sens. 51 (2013) 257–272.
[15] J.C.-W. Chan, C. Huang, R. DeFries, Enhanced algorithm performance for land cover classification from remotely sensed data using bagging and boosting, IEEE Trans. Geosci. Remote Sens. 39 (3) (2001) 693–695.
[16] M. Pal, P.M. Mather, An assessment of the effectiveness of decision tree methods for land cover classification, Remote Sens. Environ. 86 (4) (2003) 554–565.
[17] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[18] J. Ham, Y.C. Chen, M.M. Crawford, J. Ghosh, Investigation of the random forest framework for classification of hyperspectral data, IEEE Trans. Geosci. Remote Sens. 43 (3) (2005) 492–501.
[19] R.L. Lawrence, S.D. Wood, R.L. Sheley, Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest), Remote Sens. Environ. 100 (3) (2006) 356–362.
[20] J.C.-W. Chan, D. Paelinckx, Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery, Remote Sens. Environ. 112 (6) (2008) 2999–3011.
[21] X. Ceamanos, B. Waske, J.A. Benediktsson, J. Chanussot, J.R. Sveinsson, Ensemble strategies for classifying hyperspectral remote sensing data, Mult. Classif. Syst. (2009) 62–71.
[22] P. Gurram, H. Kwon, Generalized optimal kernel-based ensemble learning for hyperspectral classification problems, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2011, pp. 4431–4434.
[23] Z.H. Zhou, J.X. Wu, W. Tang, Ensembling neural networks: many could be better than all, Artif. Intell. 137 (1) (2002) 239–263.
[24] Z.H. Zhou, Ensemble Methods: Foundations and Algorithms, CRC press, 2012.
[25] D.L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory 52 (4) (2006) 1289–1306.
[26] E. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inf. Theory 52 (2) (2006) 489–509.
[27] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, S. Yan, Sparse representation for computer vision and pattern recognition, Proc. IEEE 98 (6) (2010) 1031–1044.
[28] M. Elad, Sparse and Redundant Representations: From Theory To Applications In Signal And Image Processing, Springer Verlag, New York, NY, USA, 2010.
[29] A.M. Bruckstein, D.L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Rev. 51 (1) (2009) 34–81.
[30] Z. Zhang, Y. Xu, J. Yang, X. L, A survey of sparse representation: algorithms and applications, IEEE Access (2015).
[31] X. Chen, W. Pan, J. Kwok, J. Garbonell, Accelerated gradient method for multi-task sparse learning problem, in: Proceedings of the IEEE International Conference on Data Mining, 2009.
[32] Z.H. Chen, W.M. Zuo, Q.H. Hu, L. Lin, Kernel sparse representation for time series classification, Inf. Sci. 292 (20) (2015) 15–26.
[33] L. Li, X.L. Huang, J.A.K. Suykens, Sparse recovery for jointly sparse vectors with different sensing matrices, Signal Process. 108 (2015) 451–458.

Please cite this article as: E. Zhang, et al., Spectral–spatial hyperspectral image ensemble classification via joint sparse representation, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.033

[34] B. Jiang, J. Tang, B. Luo, L. Lin, Robust feature point matching with sparse model, IEEE Trans. Image Process. 23 (12) (2014) 5175–5186.
[35] Y. Chen, N.M. Nasrabadi, T.D. Tran, Hyperspectral image classification using dictionary-based sparse representation, IEEE Trans. Geosci. Remote Sens. 10 (2011) 3973–3985.
[36] Y. Chen, N.M. Nasrabadi, T.D. Tran, Hyperspectral image classification via kernel sparse representation, IEEE Trans. Geosci. Remote Sens. 51 (1) (2013) 217–231.
[37] L. Zhang, W.-D. Zhou, Sparse ensembles using weighted combination methods based on linear programming, Pattern Recognit. 44 (2011) 97–106.
[38] K.N. Ramamurthy, J.J. Thiagarajan, A. Spanias, Ensemble sparse models for image analysis, arXiv preprint – arXiv:1302.6957, 2013.
[39] P. Gurram, H. Kwon, Ensemble learning based on multiple kernel learning for hyperspectral chemical plume detection, SPIE Defense, Security and Sensing, International Society for Optics and Photonics, 2010.
[40] P. Gurram, H. Kwon, Sparse kernel-based ensemble learning with fully optimized kernel parameters for hyperspectral classification problems, IEEE Trans. Geosci. Remote Sens. 51 (2) (2013) 787–802.
[41] L. Li, R. Stolkin, L. Jiao, F. Liu, S. Wang, A compressed sensing approach for efficient ensemble learning, Pattern Recognit. 47 (2014) 3451–3465.
[42] A. Plaza, J.A. Benediktsson, J.W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, et al., Recent advances in techniques for hyperspectral image processing, Remote Sens. Environ. 113 (2009) S110–S122.
[43] R.S. Rand, D.M. Keenan, Spatially smooth partitioning of hyperspectral imagery using spectral/spatial measures of disparity, IEEE Trans. Geosci. Remote Sens. 41 (2003) 1479–1490.
[44] 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.
[45] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC press, 1984.
[46] D. Steinberg, P. Colla, CART: classification and regression trees, Top Ten Algorithms Data Min. 9 (2009) 179.
[47] S.C. Bagui, Combining pattern classifiers: methods and algorithms, Technometrics 47 (2005) 517–518.
[48] S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1) (1998) 33–61.
[49] M. Skurichina, R.P.W. Duin, Bagging, boosting and the random subspace method for linear classifiers, Pattern Anal. Appl. 5 (2) (2002) 121–135.
[50] F. Bovolo, L. Bruzzone, M. Marconcini, A novel context-sensitive SVM for classification of remote sensing images, in: Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS), 2006, pp. 2498–2501.
[51] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, J. Calpe-Maravilla, Composite kernels for hyperspectral image classification, IEEE Geosci. Remote Sens. Lett. 3 (2006) 93–97.
[52] 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. 7 (2010) 736–740.
[53] M.V. Afonso, J.M. Bioucas-Dias, M.A.T. Figueiredo, An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems, IEEE Trans. Image Process. 20 (2011) 681–695.
[54] J. Yang, Y. Zhang, Alternating direction algorithms for ℓ_1-problems in compressive sensing, SIAM J. Sci. Comput. 33 (2011) 250–278.
[55] G. Camps-Valls, T.V.B. Marsheva, D. Zhou, Semi-supervised graph-based hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 45 (10) (2007) 3044–3054.

Erlei Zhang received the B.S. degree from the School of Electrical Engineering and Automation, Henan Polytechnic University, Jiao Zuo, China, in 2010. He is currently pursuing the Ph.D. degree in Pattern Recognition and Intelligent Systems from the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an, China. His current research interests include pattern recognition, machine learning, and remote sensing image analysis.

Xiangrong Zhang received the B.S. and M.S. degrees from the School of Computer Science, Xidian University, Xi'an, China, in 1999 and 2003, respectively, and the Ph.D. degree from the School of Electronic Engineering, Xidian University, in 2006. Currently, she is a Professor in the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University, China. She has been a Visiting Scientist in Computer Science and Artificial Intelligence Laboratory, MIT since February 2015. Her research interests include pattern recognition, machine learning, and image analysis and understanding.

Licheng Jiao received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 1982 and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively. He is the author or coauthor of more than 150 scientific papers. His current research interests include signal and image processing, learning theory and algorithms, optimization problems, and data mining.

Lin Li received the B.Sc. degree in Information and Computation Science from Xidian University, Xi'an, China, in 2009. She received her Ph.D. degree in Electronic Engineering at Xidian University in 2015. Between 2011 and 2012, she was an exchange Ph.D. student with the Centre of Excellence for Research in Computational Intelligence and Applications (CERCIA), School of Computer Science, University of Birmingham, UK. She is a Research Lecturer in the Key Laboratory of Information Fusion Technology, Ministry of Education, School of Automation, Northwestern Polytechnical University, China from May 2015. Her current research interests include computational intelligence, machine learning, numerical optimization, information fusion and image processing.

Biao Hou received the B.S. and M.S. degrees in Mathematics from Northwest University, Xi'an, China, in 1996 and 1999, respectively, and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, in 2003. Since 2003, he has been with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, where he is currently a Professor. His research interests include compressive sensing and synthetic aperture radar image interpretation, etc.

Please cite this article as: E. Zhang, et al., Spectral–spatial hyperspectral image ensemble classification via joint sparse representation, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.033