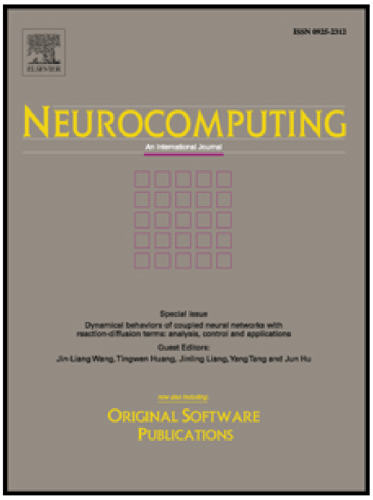
majianglin2003@gmail.com(M.J.)



ISSN 0925-2312

ELSEVIER

NEUROCOMPUTING

An International Journal

Special issue

Dynamical behaviors of coupled neural networks with reaction-diffusion terms: analysis, control and applications

Guest Editors:

Jin-Liang Wang, Tingwen Huang, Jinling Liang, Yang Tang and Jun Hu

now also including

ORIGINAL SOFTWARE PUBLICATIONS

Hyperspectral Image Classification by AdaBoost Weighted Composite Kernel Extreme Learning Machines

Lu Li , Chengyi Wang , Wei Li , Jingbo Chen

PII: S0925-2312(17)31475-3
DOI: 10.1016/j.neucom.2017.09.004
Reference: NEUCOM 18846

To appear in: *Neurocomputing*

Received date: 11 November 2016
Revised date: 15 August 2017
Accepted date: 3 September 2017

Please cite this article as: Lu Li , Chengyi Wang , Wei Li , Jingbo Chen , Hyperspectral Image Classification by AdaBoost Weighted Composite Kernel Extreme Learning Machines, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.09.004

Hyperspectral Image Classification by AdaBoost Weighted Composite Kernel Extreme Learning Machines

Lu Li [1,2,3], Chengyi Wang [2,] *, Wei Li[1] and Jingbo Chen[2]

[1] the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China; liwei089@ ieee.org(W.L.)

[2] Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Datun Road North 20A,Beijing 100101, China; lilu@radi.ac.cn(L.L.); wangcycastle@163.com(C.W.); chenjb@radi.ac.cn(J.C.); majianglin2003@gmail.com(M.J.)

[3] the University of Chinese Academy of Sciences, Beijing 100049,China ;

* Correspondence: wangcycastle@163.com; Tel.: +86-170-9008-6991

Academic Editor: name
Received: date; Accepted: date; Published: date

Abstract: Extreme learning machine (ELM) is an efficient learning algorithm for multi-classification and regression. However, original ELM doesn't consider the weight of each sample in training-set, which may cause the accuracy decreasing especially in imbalanced datasets. Even if each training sample is assigned with an extra weight, the problem on how to determinate the weight adaptively still remains. Inspiration by AdaBoost algorithm, we embed the weighted ELM algorithm in AdaBoost framework. In the meanwhile, we incorporate spatial and spectral information in composite kernel for each sample, which has a good performance in hyperspectral image (HSI) classification. By combining composite kernel methods and Adaboost framework with weighted ELM, a novel algorithm, namely AdaBoost composite kernel extreme learning machines denoted as AdaBoost-WCKELM is proposed. Experimental results demonstrate that the proposed method outperforms current state-of-the-art algorithms and derives a good improvement in HSI classification accuracy.

Keywords: Hyperspectral image Classification; Extreme learning machine; Composite Kernel; AdaBoost

PACS: J0101

1. Introduction

Hyperspectral imagery (HSI) consists of hundreds of spectral bands, hence how to extract vast amounts of information from them is the challenge remote sensing should be addressed. One of the most commonly used HSI applications is the classification on the surface of the Earth. Although many machine learning algorithms have been applied to HSI classification, such as K-nearest neighbor (KNN) [1], Logistic Regression [2, 3], Artificial Neural Network (ANN) [4] etc. Among these methods, Support Vector Machines (SVM) has been proved to have the excellent performance in terms of classification accuracy, even if training-set is small-size and contains noise. However, due to the Hughes phenomenon in HSI, the direct application of machine learning method by original bands of HSI provides poor classification performance. In [5][6], band selection is used to provide discriminative information and reduce the computational burden. However, another method by Kernel tricks could map feature vector from original lower dimensional space to high dimensional space. Furthermore, kernel strategy avoids the explicit mapping that is needed to learn a linear or nonlinear function or decision boundary. Kernel strategy was applied for HSI denoising in [7] and with SVM for HSI classification in [8]. However, traditional SVM classifier with kernel method only uses the spectral signature as input features, which could be further improved by including spatial information. There are several approach to incorporate spatial information with original spectral bands . In [9] and [10], the method based Mathematical morphology and image segmentation is introduced to HSI classification. But it is very complex. In order to incorporate spatial information simply and efficiently, [11] presented a framework of composite kernel with SVM (CKSVM) for enhanced classification accuracy of HSI. Composite kernel method also has the merit of flexibility to balance between the spatial and spectral information. There are two aspects for improvement CKSVM. Firstly, mean and standard deviation is simply selected as spatial information by original CKSVM. Different spatial feature extraction is applied to modify the framework of composite kernel. Moments of neighbor region [12], extended multi-attribute profiles (EMAPs) [12, 13, 14] and contextual information based on Markov Random Fields (MRF) [14] derived from the HSI are applied. The other improvement is that some other classifiers such as multinomial logistic regression (MLR) [12, 13] and kernel collaboration representation [15] are applied instead of SVM with composite kernel.

Recently, Huang et al. [16] have proposed an efficient Single-hidden Layer Feed-forward Neural Network (SLFN) algorithm, namely Extreme learning machines (ELMs). Compared with SVM, ELM has two evident advantages that outperform SVM. On one hand, ELM randomly generates the parameters of the nodes which are independent of training samples. Therefore, user only needs to predefine a parameter about the network architecture, i.e. the number of nodes in the hidden layer. This makes the algorithm save lots of time-consuming in tuning process. On the other hands, because algorithm analytically obtains the weight of output layer instead of iterative training-process similar as SVM, it makes the learning extremely faster than SVM. Furthermore, comparing SVM using One-Against-One (OAO) strategy in Multi-classification, ELM using One-Against-All (OAA) strategy could further reduces the computational time in both training-process and testing-process[17]. Inspired by the fact that SVM with kernel method achieves a good success, ELM could also apply kernel method, i.e. KELM. In [18], ELM using RBF as kernel function was not only more efficient but also slightly better in term of accuracy than SVM. And in [19], ELM combined with composite kernel (namely CKELM) was applied in HSI. Actually, besides the improvement of classification accuracy, another advantage of embedding kernel method in ELM framework is to avoid the generation of random parameters in hidden layer.

However, original ELMs, KELM and CKELM don't consider the weight of each sample in training-set, which may cause the accuracy decreasing especially in imbalanced datasets. Unfortunately, real HSI datasets are quite often imbalanced datasets. To address this problem, a weighted-ELM was proposed in [20] in which each training sample is assigned with an extra weight to strengthen the impact of minority class while weaken the impact of majority class. Once weight is given, it would not change both in training-process and in testing- process.

As it is known, AdaBoost framework [21] can combine weak classifiers and adjust both the weights of these weak classifiers and the weights of the training samples, adaptively. Intuitively, weighted-ELM can be drawn lessons from such adaptive adjustment of weights. In this paper, inspired by AdaBoost algorithm with weighted original ELM [22], we utilize AdaBoost framework with weighted CKELM for HSI classification, namely AdaBoost weighted composite kernel extreme learning machines (named as AdaBoost-WCKELM). At each iteration of the training process, the weight of each training sample is set according to both the current error rate of classifier and the imbalance in training-set. Therefore, the proposed method is able to adaptively determine the weights in AdaBoost framework. In the meanwhile, we optimize spatial feature extraction process by the algorithm which is similar to constant time median filter (CTMF

[23]) and saves much time-consuming. The main contribution of this paper can be summarized as follows:

1. First and foremost, our AdaBoost-WCKELM builds a bridge between weighted-CKELM and AdaBoost framework, which is, to our best knowledge, the first-time application in HSI classification.
2. Another innovative contribution of this work is to utilize composited kernel method in weighted-ELM framework for HSI classification. As far as we know, weighted-ELM has not yet been applied in hyperspectral classification before. And composited kernel method, which combines spectral bands information and spatial features, has been proven to be suitable for HSI classification when SVM or MLR is applied. But it is also the first-time that composited kernel method is applied in weighted-ELM framework for HSI classification.

The remainder of the paper is organized as follows. Section II introduces the methodology of this paper. Section III gives the experimental results and discussion. Finally, conclusions are given in Section IV.

2. Methodology

2.1. ELM, Weighted-ELM and Weighted-KELM

ELM was proposed as single-hidden layer feedforward neural networks (SLEF) algorithm. The hidden neuron parameters, namely both weight connecting vectors $\mathbf{a}$ and bias b are randomly assigned and the output weights $\boldsymbol{\beta}$ can be determinated by the Moore-Penrose generalized inverse analytically. Assuming that the training set contains N samples $\mathbf{X} = \{(\mathbf{x}_i, y_i) \mid i = 1, 2, \cdots N, \mathbf{x}_i \in \mathbf{R}^d, y_i \in \{1, 2, \cdots C\}\}$, and the hidden layer has L nodes. Where d is the number of feature dimension and C is the total number of class. The output of Original ELM (OELM) can be expressed as

$$\sum_{j=1}^{L} \beta_j \, \mathrm{g}(<\mathbf{a}_j, \mathbf{x}_i> + b_j) = T_i, i = 1, 2, \cdots, N. \tag{1}$$

where T_i is 1 if the sample $\mathbf{x}_i$ belongs to this class, otherwise -1, g denotes activation function, and $<\cdot,\cdot>$ represents inner product of two vectors. According to OAA strategy for multi-classification, the above equations of N samples in C classifiers can be rewritten as matrix model by

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{T} \tag{2}$$

Thus, we could obtain the least square solution with minimal norm as a regularization term by

$$\boldsymbol{\beta} = \mathbf{H}^{\dagger}\mathbf{T} = \begin{cases} \mathbf{H}^T(\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{T} & when \quad N \le L \\ (\lambda\mathbf{I} + \mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{T} & when \quad N \ge L \end{cases} \tag{3}$$

where $\dagger$ denotes the Moore-Penrose generalized inverse. Given a new sample $\mathbf{x}$, the predication label of original ELM classifier is obtained by

$$label(\mathbf{x}) = \arg\max_i f_i(\mathbf{x}), i = 1, 2, \cdots, C \tag{4}$$

where $\mathrm{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots f_c(\mathbf{x})]$ and

$$\mathrm{f}(\mathbf{x}) = \begin{cases} \mathbf{h}(\mathbf{x})\mathbf{H}^T(\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{T} & when \quad N \le L \\ \mathbf{h}(\mathbf{x})(\lambda\mathbf{I} + \mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{T} & when \quad N \ge L \end{cases} \tag{5}$$

OELM treats the weight of each sample in training-set equally. Actually, if $\mathbf{x}_i$ comes from a minority classes or a key sample to improve accuracy of classifier, the weight of $\mathbf{x}_i$ is larger relatively. Thus a $N \times N$ diagonal matrix $\mathbf{W}$ associated with weight w_{ii} of training sample $\mathbf{x}_i$ is considered in the weighted least square solution. And in weighted-ELM, the equation (3) can be revised as

$$\boldsymbol{\beta} = \mathbf{H}^{\dagger}\mathbf{T} = \begin{cases} \mathbf{H}^T(\lambda\mathbf{I} + \mathbf{WHH}^{\mathbf{T}})^{\mathbf{-1}}\mathbf{WT} & when \quad N \le L \\ (\lambda\mathbf{I} + \mathbf{H}^T\mathbf{WH})^{\mathbf{-1}}\mathbf{WH}^T\mathbf{T} & when \quad N \ge L \end{cases} \tag{6}$$

Inspiring by the great success of the kernel method in SVMs, kernel method could be also adopted by weighted-ELM. A kernel matrix can be defined by

$$\boldsymbol{\Omega}_{\text{ELM}} = \mathbf{HH}^{\mathbf{T}}, \boldsymbol{\Omega}_{\text{ELM}_{q,t}} = h(\mathbf{x}_q)h(\mathbf{x}_t)\text{=K}(\mathbf{x}_q, \mathbf{x}_t) \tag{7}$$

In training-process of weighted kernel ELM, the number of hidden Layer L could be regarded as just equal to the number of samples in training-set N . Therefore, we can only revise the equation (6) and (5) respectively when $N \le L$ by

$$\boldsymbol{\beta} = \mathbf{H}^{\dagger}\mathbf{T} = \mathbf{H}^T(\lambda\mathbf{I} + \mathbf{W\Omega})^{\mathbf{-1}}\mathbf{WT} \tag{8}$$

$$\text{f}(\mathbf{x})\text{=}\left[K(\mathbf{x},\mathbf{x}_1),\cdots,K(\mathbf{x},\mathbf{x}_N)\right](\lambda\mathbf{I} + \mathbf{W\Omega})^{\mathbf{-1}}\mathbf{WT} \tag{9}$$

2.2. Composited Kernel

As mentioned in [11], composite kernels could efficiently combines contextual and spectral information and have demonstrated excellent performance in HSI classification in term of accuracy and robustness. Therefore, the composite kernel, i.e. weighted summation kernel, is adopted as kernel function in our AdaBoost-WKELM. This kernel method balances the spatial and spectral content by

$$\text{K}(\mathbf{x}_i, \mathbf{x}_j) = \mu\text{K}_s(\mathbf{x}_i^s, \mathbf{x}_j^s) + (1-\mu)\text{K}_{\omega}(\mathbf{x}_i^{\omega}, \mathbf{x}_j^{\omega}) \tag{10}$$

where $\mathbf{x}^{\omega}$ and $\mathbf{x}_i^s$ are the vectors extraction by spectral band and by spatial feature from its surrounding area respectively. μ is the balance coefficient between spatial and spectral kernel method, which is [0,1]. In this paper, mean and standard deviation would be applied in a local $(2r+1)\times(2r+1)$ window per spectral band (r is the window radius).

Radial basis function (RBF) is applied for spatial kernel by

$$\text{K}_s(\mathbf{x}_i^s, \mathbf{x}_j^s)\text{=}\exp(-\gamma\left\|\mathbf{x}_i^s - \mathbf{x}_j^s\right\|^2), \gamma \in \mathbf{R}^+ \tag{11}$$

and polynomial function for spectral kernel by

$$\text{K}_{\omega}(\mathbf{x}_i^{\omega}, \mathbf{x}_j^{\omega})\text{=}\ (<\mathbf{x}_i^{\omega}, \mathbf{x}_j^{\omega}>+1)^d, d \in \mathbf{Z}^+ \tag{12}$$

2.3. AdaBoost Weighted-CKELM

AdaBoost algorithm can combine some weak classifiers to generate a single strong classifier adaptively according to the accuracy by these weak classifiers. The training step is a serial iteration. Before iteration, each sample is transformed into a new feature vector by means of user-predefined kernel functions. And the weight of each sample could be adjusted according to the performance of the classifiers in previous iteration. If a sample is misclassified by previous classifier, the weight of the sample will be increased, which makes the sample play a more important role in this iteration and forces the classifier to concentrate on the misclassified samples.

In training- process, at first, initial weight for each sample in training-set is assigned by,

$$w_{ii}^{0}(\mathbf{x}_i)=\frac{1}{C\#t_i} \tag{13}$$

where $\#t_i$ is the number of samples belonging to class t_i. The reason why we don't apply original AdaBoost using uniform weight for every sample is that imbalance usually happens in HSI. Through defined by (13), weight of each sample could be adaptively adjusted according to the imbalanced number of samples in each class. And the summation of weights of per-class is promised the same in the updated procedure.

Then, we start T iterations procedure of AdaBoost framework. At each iteration, training and constructing classifier by weighted-CKELM would be conducted just as mentioned above. And then, the accuracy of the classifier is computed and the weight of each sample in training-set is updated by

$$w_{ii}^{t+1}=\frac{w_{ii}^{t}\exp(-\alpha_t\,\mathrm{I}(label^t(\mathbf{x}_i),y_i))}{Z_{y_i}^{t}} \tag{14}$$

where α_t denotes weight of classifier and it is defined by

$$\alpha_t=\ln((1-\varepsilon_t)/\varepsilon_t)+\ln(C-1) \tag{15}$$

Notice that, the first half of (15) is just original binary-class classification in AdaBoost Framework. Due to OAA strategy applied in multi-classification, Freund and Schapire [24] extended the original AdaBoost to multiclass condition, which is successfully applied in [25].

ε_t denotes error rate of the classifiers, which is defined by

$$\varepsilon_t=\sum_i w_{ii}^{t}(1-\mathbf{I}(label^t(\mathbf{x}_i),y_i)) \tag{16}$$

$\mathrm{I}(\cdot,\cdot)$ is an indicator function

$$\mathrm{I}(a,b)=\begin{cases}1 & when\quad a=b\\ 0 & when\quad a\neq b\end{cases} \tag{17}$$

and $Z_{y_i}^{t}$ is a normalization denominator so that

$$\sum_{y_i=j} w_{ii}^{t+1}(\mathbf{x}_i)=\frac{1}{C} \tag{18}$$

(15) promises that each class plays the same important role in training- process in order to avoid classifier is inclined the class have a large number of samples in training-set.

Finally, in the testing-process, given a new sample, the label is determined by Adaboost-KELM according to the weight voted strategy,

$$label(\mathbf{x})=\arg\max_i\sum_{t=1}^{T}\alpha_t\,\mathrm{I}(label^t(\mathbf{x}),i),\ i=1,2,\cdots,C \tag{19}$$

The pseudo-code of Adaboost-KELM is shown in Algorithm 1 for more details. By compared with CKELM, the proposed AdaBoost-WCKELM has two advantages. Firstly, because it usually happens in HSI datasets that the number of samples in different class varies widely, and each training sample impacts classification with different importance. Weights are introduced in the proposed framework to rebalance the importance of each sample in per-class. It is very useful to improve classification accuracy when applied in imbalance datasets. Secondly, by AdaBoost framework, the weight of each sample could be adaptively adjusted according to both the influence by imbalance of training-samples and the current error rate of classifier. The numerator of (14) makes the weight of the training-sample larger if the sample is incorrectly classified. The denominator of (14) rearranges weight of each training-sample to rebalance the importance of each class. Therefore, when a new sample comes, outputs by series of classifiers generated by AdaBoost framework would be merged as final decision. The final decision is expected to be more robust than CKELM because only one classifier without imbalance considered is applied in CKELM. Furthermore, in framework of AdaBoost-WCKELM, once the kernel matrix of the training-set is obtained, it is reused at each iteration. Because ELM has a very high efficiency, AdaBoost framework are not much more time-consuming than CKELM.

Algorithm 1 AdaBoost-WCKELM

```
Input: Training-set $\mathbf{X} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbf{R}^d, y_i \in \{1, 2, \cdots C\}\}_{i=1}^{N}$;
       Number of iterations $T$ and regularization coefficient $\lambda$.
Initialization: $\mathbf{\Omega}_{\mathrm{ELM}}$ obtained by predefined composite kernel function by (10);
                Initial weight $w_{ii}^0$ obtained by (13).

Training- process:
for $t = 0,1,\ldots,T$-1
    $\mathbf{W}^t = \begin{bmatrix} w_{11}^t & 0 & 0 & 0 \\ 0 & w_{22}^t & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & w_{NN}^t \end{bmatrix}$;
    Training weighted-CKELM classifier with $\mathbf{W}^t$ by (8);
    Computing f($\mathbf{x}$) for each sample by (4);
    Computing label for each sample in training-set by (9);
    Computing $\alpha_t$, $\varepsilon_t$ by (15) and (16)respectively.
    If $\varepsilon_t \geq 0.5$
        $T = t - 1$ and Break;
    end
    Updating weights for each class by (14) and (18;
end
Testing-process:
Given a new sample, Computing f($\mathbf{x}$) by (9) ;
Computing label of each classifiers by (4);
Fusion the decision of each classifier by (19).
```

2.4. Fast Spatial Feature Extraction Tricks

In order to compute the mean and standard deviation efficiently, we implement a fast algorithm which is similar to CTMF. Time-consume is not relate to the size of local window and time-complexity is only O(N), N is the number of pixels in HSI.

At the beginning, we keep two values in memory. One is that summation of the intensity within local $(2r+1)\times(2r+1)$ window $\sum X$ (X represents all the data within the window) and the other one is summation of the square of the intensity $\sum X^2$. And the mean of the top-left corner pixel in HSI could be obtained by

$$M=(\sum X)/(2r+1)\times(2r+1) \tag{20}$$

And then standard deviation could be obtained by

$$\sigma=\sqrt{(\sum X^2-2M\sum X+M^2)/((2r+1)\times(2r+1)-1)} \tag{21}$$

When local window slides, as shown in Figure 1, a new calculation of $\sum X$ and $\sum X^2$ is obtained by subtract $(2r+1)$ values (red rectangle) and plus $(2r+1)$ values (blue rectangle). And the new mean and standard deviation are obtained by (20) and (21). Finally, when loops finished, mean and standard deviation of all pixels in HSI are obtained, which represents the input of spatial kernel function.

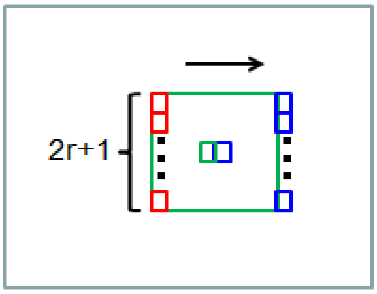


Figure 1. Fast mean and standard deviation obtained.

3. Experimental Result and Discussion

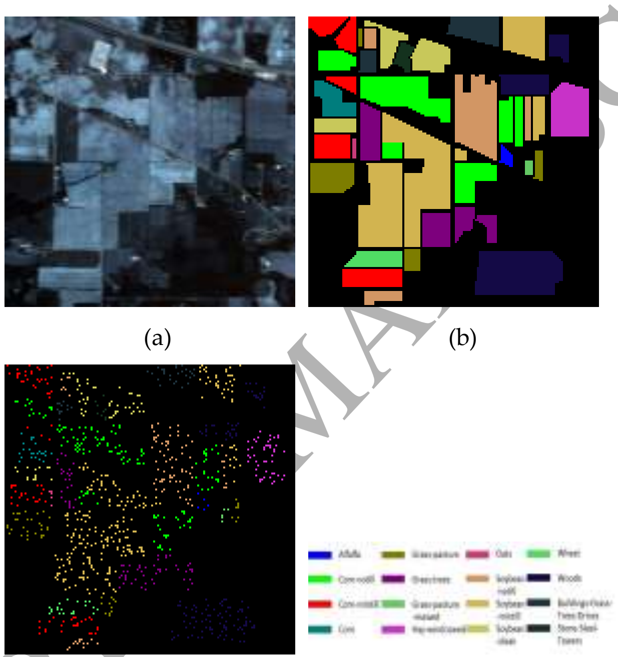
In this section, we evaluate our proposed method by two real hyperspectral datasets[1]. The environment of all experiments is a desktop PC equipped with an Intel(R) Xeon(R) CPU E5-2620 and 24GB of RAM. The implement code is available when contracting author.

3.1. Hyperspectral Datasets and Training-sets

1) The first HSI is a 2×2 mile portion of agricultural area over the Indian Pines region

[1] The three HSI benchmarks are available from: http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

in Northwest Indiana, which was acquired by NASA's Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor. This scene with a size of 145×145 pixels, comprises 202 spectral bands in the wavelength range from 0.4to 2.5 μm, with spatial resolution of 20m. The ground truth of scene contains 16 classes of interest and total 10366 samples. Due to the imbalanced number of available labeled pixels and a large number of mixed pixels per class, this dataset give a challenge to HSI classification. In order to evaluate our proposed method, we only choose 10% samples per- class to our training-set. Because of five-fold cross validation for parameter tuning, if the number of samples in any class is smaller than five, we choose five samples to training-set. The rest samples consist of testing-set. Figure 2 shows a false-color image, ground truth and training –set of Indian Pines dataset. Table 1 presents the class descriptions and sample distributions for both training-set and testing-set.



(c) (d)

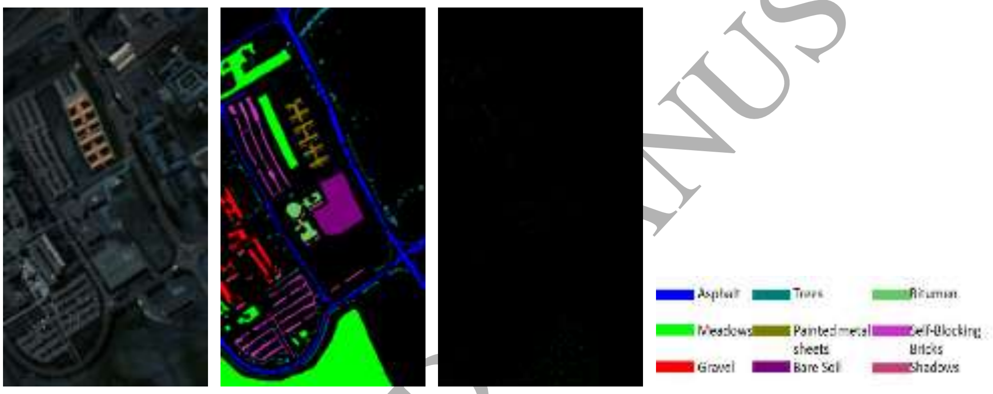
Figure 2. Indian Pines dataset (a) False color image of the Indian Pines image by bands 10, 20 and 30 for red, green and blue respectively; (b) Ground truth data of the Indian Pines image; (c) Samples in training-set and (d) Reference map.

Table 1. Per-class samples for training-set and testing-set in Indiana-Pines Dataset

class		Numbers of Samples	
No.	**Name**	**Training-set**	**Testing-set**
1	ALFALFA	5	41
2	CORN-N	143	1285
3	CORN-M	83	747
4	CORN	24	213
5	GRASS-PASTURE	48	435
6	GRASS-TREES	73	657
7	GRASS-P-M	5	23
8	HAY-W	48	430

9	OATS	5	15
10	SOYBEAN-N	97	875
11	SOYBEAN-M	246	2209
12	SOYBEAN-C	59	534
13	WHEAT	21	184
14	WOODS	127	1138
15	BUILDINGS-G-T-D	39	347
16	STONE-S-T	9	84

2) The second HSI is a 103-bands image acquired by Reflective Optics Spectrographic Image System (ROSIS-03) sensor over the urban area of the University of Pavia, Italy. The spatial resolution is 1.3m and the scene contains 610×340 pixels and 9 classes. The number of samples is 42776 in total. We choose 1% samples per-class to evaluate our proposed method. Figure 3 shows a false-color image ground truth and training-set of Pavia University dataset.



(a) (b) (c) (d)

Figure 3. Pavia University dataset (a) False color image of the Indian Pines image by bands 20, 40 and 60 for red, green and blue respectively; (b) Ground truth data of the Indian Pines image; (c) Samples in training-set and (d) Reference map.

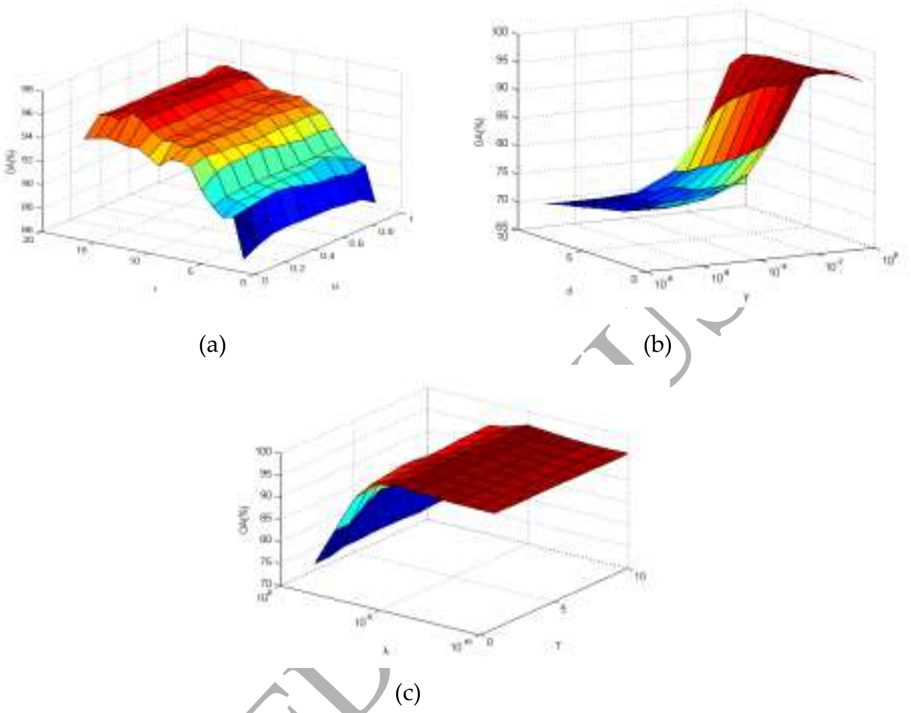
Table 2. Per-class samples for training-set and testing-set in pavia university dataset

class		Numbers of Samples	
No.	**Name**	**Training-set**	**Testing-set**
1	ASPHALT	66	6565
2	MEADOWS	186	18463
3	GRAVEL	21	2078
4	TREES	31	3033
5	PAINTED-M-S	13	1332
6	BARE SOIL	50	4979
7	BITUMEN	13	1317
8	SELF-B B	37	3645
9	SHADOWS	9	938

3.2 Parameters Tuning

First of all, we study the parameters of AdaBoost-WCKELM for HSI classification. There are two parameters for AdaBoost framework, i.e., number of iterations T and regularization coefficient λ, three parameters for kernel function, i.e., μ,γ,d, and the radius r of local window for spatial feature extraction. Five-fold cross validation is used to tune parameters by training-set in terms of Overall-Accuracy (OA). Parameter searching

scale $r \in \{1,2,\cdots,20\}$, $\mu \in \{0,0.1,\cdots,0.9\}$, $\gamma \in \{10^{0},10^{-1},\cdots 10^{-7}\}$, $d \in \{1,2,\cdots 10\}$, $\lambda \in \{10^{-1},10^{-2},\cdots,10^{-10}\}$ and $T \in \{1,2,\cdots,10\}$, respectively.



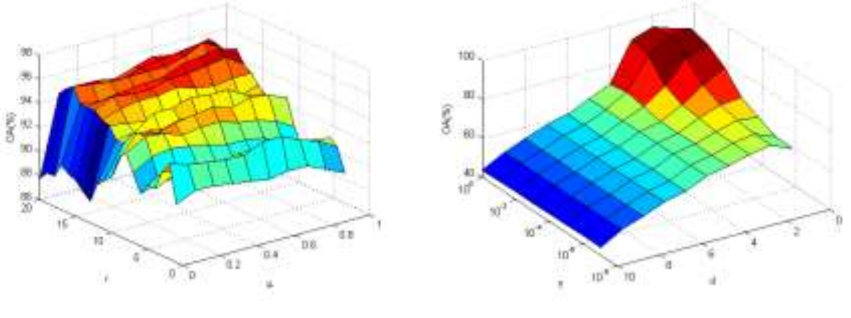
(a)

(b)

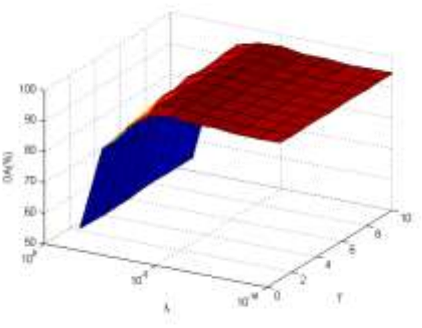
(c)

Figure 4. Three groups of parameters optimization by grid search method in Indian Pines dataset. (a) $\{\mu,r\}$; (b) $\{\gamma,d\}$; and(c) $\{\lambda,T\}$.

In order to improve the search efficiency, we divide the parameters into three groups, i.e. $\{\mu,r\}$, $\{\gamma,d\}$ and $\{\lambda,T\}$.And then two-dimensional grid search is performed in each group alternately until there is no improvement of OA. Figure 4 and Figure 5 show the three groups of parameters optimization by grid search method in the Indian Pines dataset and the Pavia University dataset, respectively. Finally we set $\{\mathrm{r},\mu,\gamma,d,\lambda,T\}$ to $\{13,0.3,0.1,1,10^{-5},6\}$ for the Indian Pines datasets and $\{18,0.8,0.1,1,10^{-5},5\}$ for the Pavia University respectively.



(a) (b)



(c)

Figure 5. Three groups of parameters optimization by grid search method in Pavia University dataset. (a) $\{\mu,r\}$;(b) $\{\gamma,d\}$; and(c) $\{\lambda,T\}$.

3.3 Experimental Results and Analysis

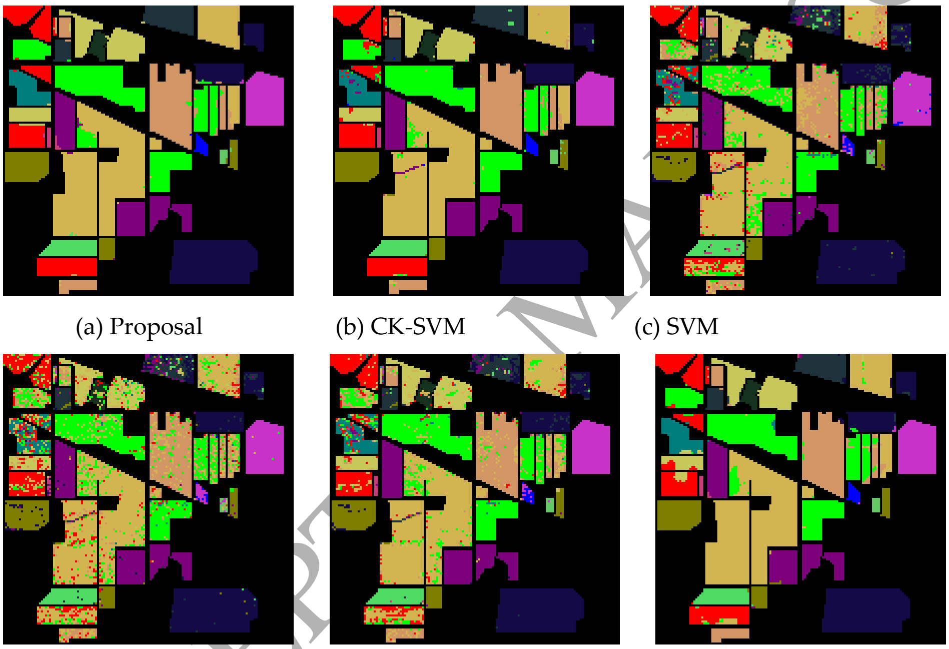
In this section, the proposed method is evaluated and compared with several widely used classification methods including SVM, ELM, KELM, CKSVM, CKELM in terms of OAs, Average-Accuracy (AA) and κ statistic in both HSI datasets. The RBF kernel has been applied in SVM and KELM. The optimal parameters C and γ are chosen by fivefold cross validation in SVM and CKSVM. The recommended values of CKSVM and CKELM parameters μ and r in [17] are adopted, Sigmoid is applied as activation function in ELM and the number of the hidden layers in ELM and the regularization coefficient λ in KELM is also chosen by fivefold cross validation.

We test the performance of our proposed method in Indiana Pines dataset. The classification accuracy including OA, AA, κ statistic and individual class accuracies are reported in Table 3. And Figure 6 shows some classification maps by different methods. Proposed method obtains OA of 98.08%, AA of 97.67% and κ of 97.81%, which are the best classification results in our experiments. Furthermore, the proposed method achieves 11 the highest individual class accuracies among 16 classes in all, especially the OA of the class containing very small training samples, e.g. Oats and Alfalfa, obtains 100%. Therefore, the proposed method is suitable for imbalanced training-set. Then the same test procedure is conducted in the Pavia University dataset. The classification evaluation is reported in Table 4 and classification maps are shown in Figure 7. The proposed method obtains OA of 96.46%, AA of 93.32% and κ of 95.31%, which are also the best classification results in our experiments. There are 5 classes obtained the highest individual class accuracies among 9 classes.

TABLE 3. Indian Pines classification accuracy(%) (The Best result Are Highlighted in Bold Typeface)

class	SVM	ELM	KELM	CKSVM	CKELM	Proposal
Alfalfa	56.10	12.20	58.53	97.56	97.56	**100**
Corn-N	73.15	70.04	82.80	91.75	93.54	**97.74**
Corn-M	70.68	52.21	63.99	96.65	90.09	**97.59**
Corn	67.14	40.85	65.26	92.02	92.02	**98.12**
Grass-P	87.36	85.75	90.11	94.48	95.17	**95.17**

Grass-T	97.26	97.26	99.39	**100**	97.87	98.32
Grass-P-M	86.96	47.83	95.65	**100**	**100**	86.96
Hay-W	96.74	99.07	100	**100**	**100**	**100**
Oats	93.33	40.00	93.33	**100**	**100**	**100**
Soybean-N	76.57	67.54	80.23	97.94	96.68	**98.17**
Soybean-M	76.19	73.91	83.30	96.60	97.55	**97.87**
Soybean-C	83.90	63.67	84.83	96.25	92.13	**99.81**
Wheat	90.22	96.74	97.28	99.45	98.37	**100**
Woods	91.74	96.13	95.34	**99.65**	98.33	98.51
B-G-T-D	62.25	49.57	68.01	97.12	96.54	**97.98**
Stone-S-T	90.48	59.52	84.52	**98.81**	**98.81**	96.43
OA	80.35	74.66	84.43	96.72	95.99	**98.08**
AA	81.25	65.73	83.91	97.39	96.54	**97.67**
κ	77.60	70.99	82.20	96.27	95.42	**97.81**



(a) Proposal (b) CK-SVM (c) SVM

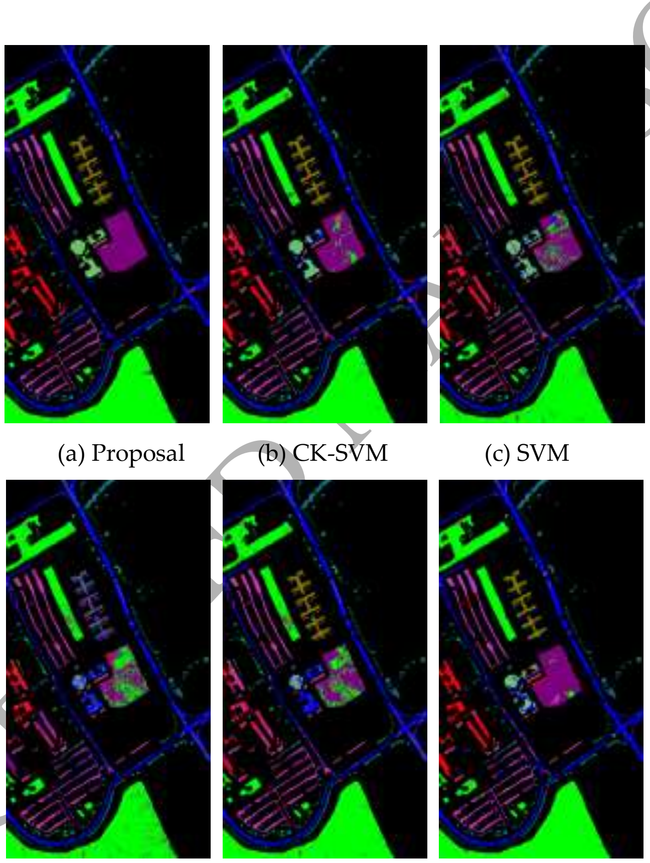
(d) ELM (e) KELM (f) CKELM

Figure 6. Classification maps obtained for the Indiana Pines dataset by (a)Proposal ,(b)CKSVM,(c)SVM,(d)ELM,(e)KELM,(f)CKELM.

By comparison of different methods in detail, we find that the accuracy obtained by SVM is better than that obtained by ELM, especially when the number of training samples is very small. By adopting the kernel method, the classification accuracy by KELM can be improved. Due to spatial information incorporating with spectral information, CKSVM and CKELM obtain the better result than SVM and KELM. Because the proposed method combines Adaboost framework with CKELM, joint-decision is made by different classifiers for HSI datasets, which makes the proposed method more robust than traditional CKSVM and CKELM.

TABLE 4. Pavia University classification accuracy(%) (The Best result Are Highlighted in Bold Typeface)

class	SVM	ELM	KELM	CKSVM	CKELM	Proposal
Asphalt	88.19	90.59	88.96	90.43	93.88	**97.29**
Meadows	96.73	95.40	97.49	98.39	**98.88**	98.48
Gravel	75.02	54.81	50.67	82.10	92.97	**93.74**
Trees	92.45	78.37	87.67	**95.09**	91.03	92.18
Painted-M-S	99.02	1.2	98.20	99.62	**100**	87.82
Bare Soil	77.61	54.07	67.89	88.57	93.53	**99.50**
Bitumen	76.99	26.27	24.75	74.33	61.20	**96.58**
Self-B B	82.28	75.22	88.45	90.29	89.11	**91.17**
Shadows	**99.57**	97.65	97.76	96.37	66.95	73.13
OA	90.06	79.79	86.66	93.51	93.94	**96.46**
AA	87.54	63.73	77.98	90.58	87.51	**93.32**
κ	86.75	72.47	81.97	91.37	91.95	**95.31**


(a) Proposal (b) CK-SVM (c) SVM

(d) ELM (e) KELM (f) CKELM

Figure 7. Classification maps obtained for the Pavia University dataset by (a)Proposal ,(b)CKSVM,(c)SVM,(d)ELM,(e)KELM,(f)CKELM.

3.4 Influence by different sizes of training-set

Additionally, we analyze the sensitivity of the different sizes of training-set. For this purpose, we further design different training-sets of the Pavia University dataset by randomly choosing 20-200 labeled samples per-class. We also evaluate the performance of several classifiers, such as CKSVM, SVM, ELM and KELM. Furthermore, we add CKELM for comparison, which just adopts ELM with composite kernel method but without

AdaBoost framework. Therefore, CKELM can be realized just by setting T=1 in AdaBoost-WCKELM. The classification results are shown in Figure 8.

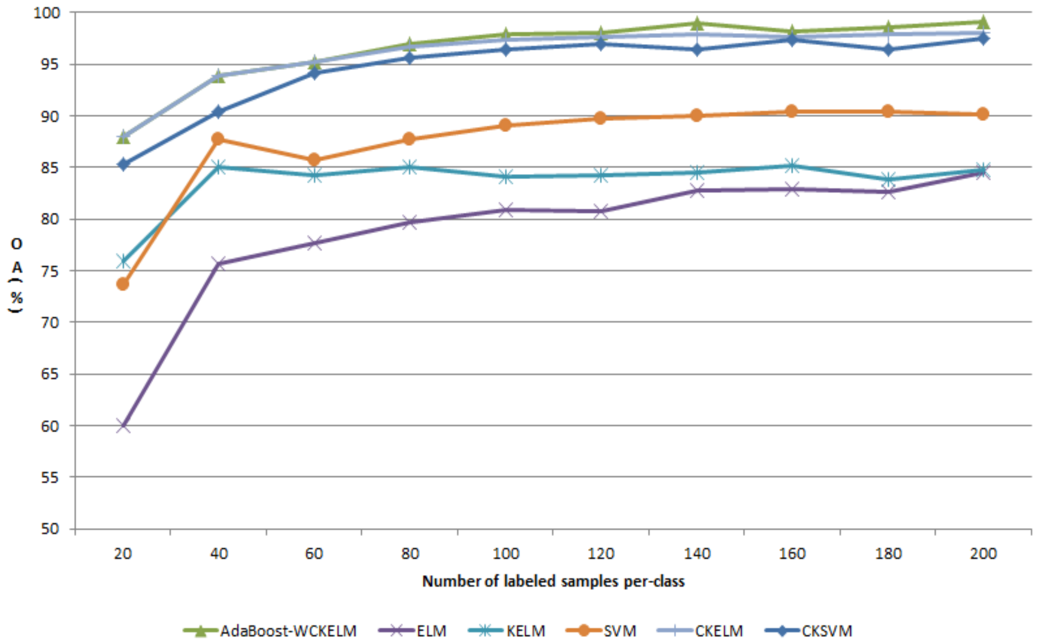


Figure 8. OA results by different methods as a function of different numbers of labeled samples per class for Pavia University dataset.

From Figure 8, the proposed method provides the highest accuracy regardless of different sizes of the training-set. OA of CKELM is very close to that of the proposed method when the size of the training-set is very small. The reason is that when the number of labeled samples is small, the correct rate of samples in training-set is much easier to achieve to 100% and weighted adjustment is exterminated. Thus, there are very few classifiers to be constructed. However, when the labeled samples increase, there are more classifiers to be generated in training process. The experimental results in Figure 8, (OA of the proposed method is 1.08% higher than that of CKELM when 140 samples in per-class are chosen) are proved the effectiveness of the proposed method. Another significantly observation we can derive from Figure 8 is that the methods with spatial information and spectral information are much better than those only with spectral information.

4. Conclusions

In this paper, a novel HSI classification method was proposed, i.e., AdaBoost-WCKELM. The proposed method combines ELM algorithm with AdaBoost framework to adjust weights of training samples adaptively. In order to faster embed spatial information with original spectral bands, local mean and standard have been extracted. Then, composite kernel method was applied in ELM algorithm. The experimental results demonstrated that the proposed method could provide good accuracies in both HSI dataset even if the training samples are imbalanced or very limited. It also exhibited robustness and excellent performance to different sizes of training-set when compared with other methods.

Although the proposed method is competitive with other state-of-the-art methods, such as CKSVM, there are still two important research directions deserving future attention. On one hand, the aggregation of different type machine learning classifiers in our AdaBoost framework could be considered to improve classification accuracy. On the other hand, we just utilize simple spatial feature, i.e., local mean and standard. In the future, some other spatial feature extraction, e.g., Gabor filter, morphological feature, etc., would be embedding sequential in AdaBoost-WCKELM.

References

1. Kuo C.; Yang M.; Sheu W.; et al., Kernel-Based KNN and Gaussian Classifiers for Hyperspectral Image Classification. Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International IEEE, 2008, II-1006 - II-1008.
2. D. Böhning; Multinomial logistic regression algorithm,.Ann. Inst. Stat. Math., 1992, 44,197–200.
3. J. Li; J. Bioucas-Dias; A. Plaza; Semi-supervised hyperspectral image segmentation using multinomial logistic regression with active learning. IEEE Trans. Geoscience & Remote Sensing , 2010, 48,4085–4098.
4. Ratle F.; Camps-Valls, G.; Weston J.; Semisupervised neural networks for efficient hyperspectral image classification. IEEE Trans. Geoscience & Remote Sensing, 2010, 48, 2271–2282.
5. Y.Yuan; X. Zheng; X. Lu; Discovering Diverse Subset for Unsupervised Hyperspectral Band Selection. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society. 2016,26,51-64.
6. X. Zheng; Y.Yuan; X. Lu; Dimensionality reduction by spatial-spectral preservation in selected bands. IEEE Transactions on Geoscience & Remote Sensing, 2017, 99, 1-13.
7. Y. Yuan; X. Zheng; X. Lu.; Spectral–Spatial Kernel Regularized for Hyperspectral Image Denoising. IEEE Transactions on Geoscience & Remote Sensing, 2015, 53, 3815-3832.
8. Melgani, F.; Bruzzone, L.; Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans. Geoscience & Remote Sensing. 2004, 42, 1778–1790.
9. F. Mathieu; Y. Tarabalka; J. Benediktsson and et al.; Advances in Spectral-Spatial Classification of Hyperspectral Images. Proceedings of the IEEE 101.2013, 3,652-675.
10. A. Erchan; Hyperspectral Image Classification With Multidimensional Attribute Profiles. IEEE Geoscience & Remote Sensing Letters. 2015,12,2031-2035.
11. Camps V. G.; Gomez C. L.; Munoz M. J.; et al., Composite kernels for hyperspectral image classification. IEEE Geoscience & Remote Sensing Letters, 2006, 3, 93-97.
12. Rafika S.; Karim E.; Mohamed H.; Spectral-spatial classification of hyperspectral images using different spatial features and composite kernels. Image Processing, Applications and Systems Conference (IPAS), 2014 , 1-7. Online Available: http://dx.doi.org/10.1109/IPAS.2014.7043323.
13. Li J; Reddy P.; Plaza A.; et al., Generalized Composite Kernel Framework for Hyperspectral Image Classification. IEEE Trans. Geoscience & Remote Sensing, 2013, 51,4816-4829.
14. Jianjun L.; Xiaoqian S.; Zebin W.; et al., Hyperspectral Image classification via Region-based Composite Kernels. 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS),2016, 933-936. Online Available: http://dx.doi.org/10.1109/IGARSS.2016.7729236.
15. I-Ling C.; Kai-Chih P.; Jinn-Min Y.; et al., An automatic method to determine the coefficient of the composite kernel for hyperspectral image classification. Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International, 2011, 1704-1707. Online Available: http://dx.doi.org/10.1109/IGARSS.2011.6049563
16. Huang G.B.; Zhu Q.Y.; Siew C.-K; Extreme learning machine: Theory and applications. Neurocomputing, 2006, 70, 489–501.
17. Huang G.-B.; Zhou, H.; Ding, X.; et al., Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, 2012, 42,513-29.
18. Pal M.; Maxwell A. E.; Warner T. A.; Kernel-based extreme learning machine for remote-sensing image classification. Remote Sensing Letters, 2013, 4, 853–862.

19. Yicong Z.; Jiangtao P.; Philip C.; Extreme Learning Machine with Composite Kernels for Hyperspectral Image Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015, 8, 2351-236.
20. W. Zong; G.-B. Huang; Y. Chen; "Weighted extreme learning machine for imbalance learning", Neurocomputing, 2013,101, pp.229–242.
21. Y.Freund; Robert E.S.; A short introduction to Boosting. Journal of Japenese society for artificial intelligence. 1999, 14, 771-780.
22. Kuan L.; Xiangfei K.; Zhi L.; et al., Boosting weighted ELM for imblalanced learning. Neurocomputing, 2014,123, 15-31.
23. Perreault S.; Hébert P.; Median filtering in constant time. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2007, 16, 2389-94.
24. Shen Y.; Jiang Y.; Liu W., et al.; Multi-class AdaBoost ELM. Proceedings of ELM-2014, Springer International Publishing, 2015, 2. 179-188.
25. Y. Freund; R. E. Schapire; A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 1997, 55, 119–139.

Acknowledgments: This research was supported in part by the Natural Science Foundation of China with Projects ID 41301499. The authors would like to thank Prof. D. Landgrebe, Prof. P. Gamba (Pavia University dataset), and Prof. M. Crawford (Indiana Pines dataset) for making the HSI datasets and ground truth used in the experiments. The authors would also like to thank the Associate Editor and the anonymous reviewers for their constructive comments and suggestions, which greatly helped us improve the quality and presentation of this paper.

Author Contributions: Lu Li and Chengyi Wang conceived and designed the experiments; Lu Li and Wei Li performed the experiments; Lu Li and Jingbo Chen. analyzed the data; Lu Li and Wei Li wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Li Lu(S'15) received the B.Sc. and M.Sc. degree from Wuhan University, Wuhan, China, in 2003, and 2007 respectively. He received the Ph.D. degree in the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China in 2017. He is currently a Postdoctoral Researcher in the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China. His current research interests include image classification, image processing, and photogrammetry.

Chengyi Wang received his Ph.D. degree in Institute of Remote Sensing Applications Chinese Academy of Sciences, Beijing, China, in 2007. He is now an Associate Professor of Institute of Remote Sensing and Digital Earth Chinese Academy of Sciences. His research interests cover image processing, computer vision and 3D reconstruction.

Wei Li received the B.E. degree in Telecommunications Engineering from Xidian University, Xi'an, China, in 2007, the M.S. degree in Information Science and Technology from Sun Yat-sen University, Guangzhou, China, in 2009, and the Ph.D. degree in Electrical and Computer Engineering from Mississippi State University, Starkville, Mississippi, USA, in 2012. Wei Li is currently a Professor of the College of Information Science and Technolgy at Beijing University of Chemical Technology, Beijing, China. He held a Postdoctoral Research position at University of California, Davis, CA, USA.

His current research interests include digital Image Processing, feature extraction, pattern classification, anomaly detection, data reconstruction, Hyperspectral Imagery.

Jingbo Chen received his B.Sc. and M.S.c degree in Liaoning Technical University in 2005 and 2008 respectively. He received his Ph.D. degree in Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2011.He is now an Assist-ant Professor of Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China. His research interests cover land-use classification and change detection, deep learning in remote sensing.