SUPERVISED CLASSIFICATION OF HYPERSPECTRAL IMAGES USING LOCAL-RECEPTIVE-FIELDS-BASED KERNEL EXTREME LEARNING MACHINE

Yu Shen[1], *Jianyu Chen*[2], *Liang Xiao*[1*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China;
[2]Second Institute of Oceanography, State Oceanic Administration, Hangzhou 310012, China

ABSTRACT

In this paper, we propose a local receptive fields (LRF) based kernel extreme learning machine (KELM) method for hyperspectral image (HSI) classification. As a single-hidden-layer feed-forward neural networks, kernel ELM has been used with success for classification of HSI. Considering local correlations in the spatial domain of HSI, the local random convolution nodes are introduced as LRF in the input layer. By using the LRF-based convolutional ELM feature learning on the principle component images, the proposed method can automatically learn rich feature hierarchies. All the learnt rich feature hierarchies are compressed to compact features, and the kernel ELM is applied to obtain the final classification results. Experimental results on the widely used real HSI data set indicate that the proposed approach outperforms several well-known classification methods.

Index Terms— hyperspectral image classification, LRF-KELM, random convolution nodes, SVM

1. INTRODUCTION

Recent developments in remote sensing technologies allow us to collect enormous of high spectral resolution hyperspectral images (HSIs) via airbones and spaceborne platforms. Due to the abundant spectral information in HSI bands, supervised classification using HSI data has attracted increasing interests and become a hot research area over the past decades.

For HSI classification, many machine learning methods have been developed, such as support vector machine (SVM) [1], multinomial logistic regression (MLR) [2] and extreme learning machine (ELM) [3]. Among these supervised HSI machine learning techniques, ELM has been used with success. In particular, ELM randomly generates the hidden node parameters and analytically determines the output weights instead of iterative tuning, which makes the learning extremely fast. ELM is not only computationally efficient but also tends to achieve similar or even better generalization performance than SVMs. In [4], ELM was used for land cover classification, which achieved comparable classification accuracies to a back-propagation neural network on two datasets considered. KELM was used in [5] for hyperspectral remote sensing images classification. The results indicate that KELM is similar to, or more accurate than, SVMs in terms of classification accuracy and offers notably low computational cost. However, pure pixel-wise ELM or KELM classification can only consider spectral information while spatial information is ignored. To cope this drawback, many pre-processing methods were proposed to obtain spatial feature such as Gabor filtering [6] and local binary pattern (LBP) [7]. These methods usually contain two separate steps: spatial-spectral feature extraction for pre-processing and ELM classification.

Recently, local receptive fields based ELM [8] is proposed for image or texture classification, which successfully integrates feature mapping and classification in one network. LRF-ELM applies LRF based random convolution as the input layer. According to ELM theories, the LRF can generate random convolution nodes according to continuous probability distribution. In [9], KELM with LRF (Spectral-LRF-KELM) was first used for HSI classification. Unfortunately, the work merely considered single spectral band while ignored the spatial-spectral information of HSI.

In this work, we extend LRF-KELM classification method from 2D image to hyperspectral data cube. First, we use principal component analysis (PCA) [10] as a dimension reduction technique to produce characteristic images of a HSI. Second, several different random input convolution operators are used on the principal components (PCs) to obtain feature maps. Then we apply pooling operation on these feature maps. After merging these feature maps, we use PCA again to reduce the high dimension of the feature cube. At last, KELM is applied to obtain the final classification results.

2. BRIEF OF ELM

As a single layer feedforward neural network, ELM [3] does not need to tune the hidden layer parameters. Compared with traditional classifiers, its computational cost is much lower.

This work has been supported by the National Natural Science Foundation of China (Grant No. 61571230), the National Major Research Plan of China (Grant No. 2016YFF0103604), the Jiangsu Provincial Natural Science Foundation of China (Grant No. BK20161500), the Jiangsu Provincial 333 Project. (*Corresponding author*: Liang Xiao.)

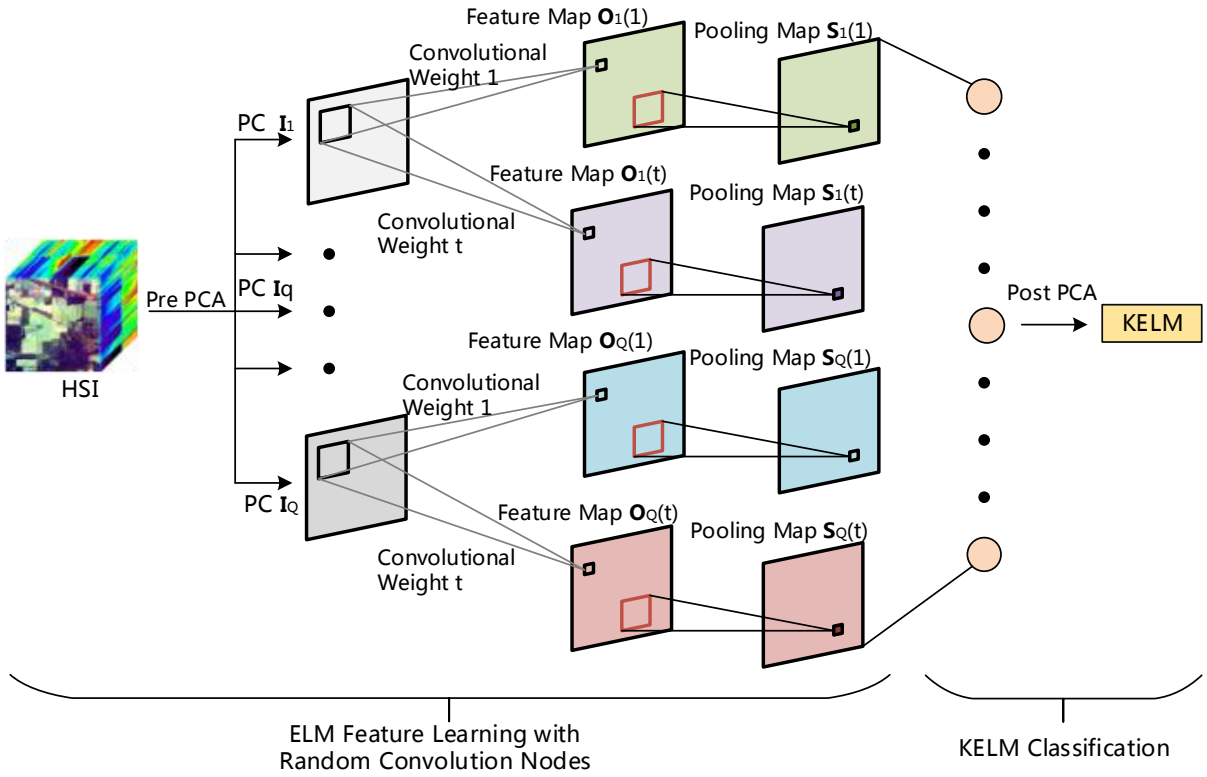


Fig. 1. A schematic of proposed LRF-KELM based HSI classification.

Given C classes of a HSI, let the class labels be defined as $z_k \in \{0, 1\}$, $1 \leq k \leq C$. Thus, a constructed row vector $\mathbf{z} = [z_1, ..., z_k, ...z_C]$ indicates the class to which a sample belongs to. For example, if $z_k = 1$ and other elements in $\mathbf{z}$ are zero, then the sample belongs to the 1st class. Given G training samples $\{\mathbf{x}_i, \mathbf{z}_i\}_{i=1}^{G}$ belonging to C classes, where $\mathbf{z}_i \in \mathbb{R}^C$ and $\mathbf{x}_i \in \mathbb{R}^B$($B$ is the number of the spectral bands or dimensionality). The output function of an ELM having L hidden neurons can be represented as

$$f(\mathbf{x}_i) = \sum_{j=1}^{L} \boldsymbol{\beta}_j h\left(\boldsymbol{\omega}_j \cdot \mathbf{x}_i + e_j\right) = \mathbf{z}_i, \tag{1}$$

where $h(\cdot)$ is a nonlinear activation function (e.g. Sigmoid function), $\boldsymbol{\beta}_j \in \mathbb{R}^c$ is the weight vector connecting the jth hidden neuron and the output neurons, $\boldsymbol{\omega}_j \in \mathbb{R}^B$ is the weight vector connecting the jth hidden neuron and input neurons, and e_j is the bias of the jth hidden neuron. $\boldsymbol{\omega}_j \cdot \mathbf{x}_i$ denotes the inner product of $\boldsymbol{\omega}_j$ and $\mathbf{x}_i$. With G equations, Equation (1) can be written compactly as

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{Z}, \tag{2}$$

where the label matrix $\mathbf{Z} = [\mathbf{z}_1^T \mathbf{z}_2^T \mathbf{z}_3^T \dots \mathbf{z}_G^T]^T \in \mathbb{R}^{G \times C}$, $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^T \boldsymbol{\beta}_2^T \boldsymbol{\beta}_3^T \dots \boldsymbol{\beta}_L^T]^T \in \mathbb{R}^{L \times C}$, and $\mathbf{H}$ is the hidden layer output matrix of the neural network. The solution of Equation (2) proposed in [3] is defined as

$$\boldsymbol{\beta}' = \mathbf{H}^{\dagger}\mathbf{Z}, \tag{3}$$

where $\mathbf{H}^{\dagger}$ is the *Moore-Penrose* generalized inverse of matrix $\mathbf{H}$. The *Moore-Penrose* generalized inverse of $\mathbf{H}$ can be calculated as $\mathbf{H}^{\dagger} = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}$. For better stability and generalization, a positive value ρ^{-1} is added to the diagonal elements of $\mathbf{H}\mathbf{H}^T$. Therefore, we have the output function of ELM classifier

$$f(\mathbf{x}_i) = \mathbf{h}(\mathbf{x}_i)\boldsymbol{\beta} = \mathbf{h}(\mathbf{x}_i)\mathrm{H}^T\left(\frac{\mathbf{I}}{\rho} + \mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{Z}. \tag{4}$$

3. HSI CLASSIFICATION VIA LRF-KELM

Fig.1 shows a schematic of the proposed LRF-KELM-based HSI classification method, which consists of two main parts: 1) LRF based feature mapping with random convolution nodes; 2) classification with KELM. By utilizing several diverse random convolutional nodes and pooling operations in PCs of HSI, local spatial structures are learned. KELM applies a kernel function for ELM and achieves better classification performance on HSI data.

3.1. LRF-based Feature Mapping

Fig.1 depicts the implementation of LRF feature mapping network, including random convolution layer and pooling layer.

Let $\mathbf{I} = [\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, \dots, \mathbf{I}_Q]$ ($\mathbf{I}_q \in \mathbb{R}^{M \times N}$)be the first Q PCs of a HSI (M and N be the size of every PC), $w_t \in \mathbb{R}^{r \times r}$

($t = 1, \ldots, T$)be the random convolution weight with T diverse feature maps, r is the window size of convolution node. As shown in Fig. 1, convolutional layer is composed of random convolutional nodes. The input convolution weights of the same feature maps are shared while distinct among different maps. The convolutional node $\mathbf{O}_q(i, j, t)$ in the tth feature map can be calculated as

$$\mathbf{O}_q(i, j, t) = \sum_{(x,y)\in\mathcal{N}(i,j)} \mathbf{I}_q(x, y)w_t, \tag{5}$$

where $\mathbf{I}_q(x, y)$ is represented as the neighboring pixels of $\mathbf{O}_q(i, j, t)$ in a convolutional window. In particular, we generate the random input weights w_t from the standard Gaussian distribution.

The pooling structure used here is average pooling. Assuming the pooling window size is $d \times d$, so the pooling node in the tth map of the q PC can be written as

$$\mathbf{S}_q(u, v, t) = \frac{1}{d \times d} \sum_{(i,j)\in\mathcal{N}(u,v)} \mathbf{O}_q(i, j, t). \tag{6}$$

Average pooling operations play an important role in the networks, which contribute data variation and perturbation. In this work, we use average pooling instead of square root pooling in [8] because both pooling operations can achieve similar results and average pooling has a simple structure.

In LRF-based feature mapping part, Q PCs are mapped with T random convolution nodes, which means $Q\times T$ feature maps are generated.

3.2. KELM Learning Classification

After LRF-based feature mapping, $Q \times T$ feature maps are merged as a new feature cube. Due to the high dimension of the feature cube, PCA is used here to reduce feature cube's dimension for computation simplicity.

To get better classification performance, we use KELM instead of ELM for classification. Let $\{\widetilde{x}\}_{i=1}^{G}$ be the samples of feature data after PCA dimension reduction. As the inner production involved $\mathbf{H}\mathbf{H}^T$ and $\mathbf{h}(\mathbf{x}_i)\mathbf{H}^T$ in Equation (4) can be replaced by a kernel function, the kernel matrix for ELM can be defined as:

$$\mathbf{\Omega}_{ELM} = \mathbf{H}\mathbf{H}^T : \mathbf{\Omega}_{ELM} = \mathbf{h}(\widetilde{\mathbf{x}}_i) \cdot \mathbf{h}(\widetilde{\mathbf{x}}_j) = \mathbf{K}(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}_j). \tag{7}$$

Thus, the output function of KELM can be written as

$$f(\widetilde{\mathbf{x}}_i) = \begin{bmatrix} \mathbf{K}(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}_1) \\ \vdots \\ \mathbf{K}(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}_G) \end{bmatrix} (\frac{\mathbf{I}}{\rho} + \mathbf{\Omega}_{ELM})^{-1}\mathbf{Z}. \tag{8}$$

The label of input data is determined by the index of the output node with the largest value.

The LRF-KELM can be regarded as a special type of ELM learning method. To cope HSI classification problem, we extend the model to high dimension situation. Different from full connections, local random convolution nodes generate local spatial features. Then we can obtain the classification results via KELM.

4. EXPERIMENTS

4.1. HSI Dataset

To validate the effectiveness of the proposed method, we adopt the commonly-used Indian Pines data set gathered by the Airborne/Visible Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines test site in northwestern Indiana. The AVIRIS sensor generates 220 spectral bands ranging from 0.2 to 2.4μm and has a spatial resolution of 20m per pixel. The hyperspectral image consists of a $145 \times 145 \times 220$ data cube reduced to 200 bands after removing the water absorption and noisy bands. The ground truth contains 16 land cover classes and 10366 labeled pixels as seen in Table 1.

In order to avoid any bias induced by random sampling, all the experiments are repeated ten times with different randomly selected training sets. Overall accuracy (OA), average accuracy (AA) and the kappa coefficient of agreement (k) are used to assess the classification performance. All experiments are carried out using MATLAB 2014a on an Intel Xeon 3.00-GHz machine with 8 GB of RAM.

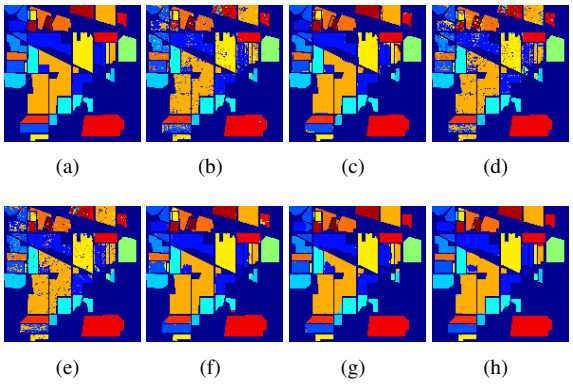
4.2. Experimental Results

The classification performance of the proposed method LRF-KELM was evaluated by comparing with KSVM [11], SVM-CK [12], KELM [5], Spectral-LRF-KELM [9], Bilateral-filtering-based KELM (BF-KELM) [13], LRF-ELM as well as LRF-KELM. The RBF kernel was adopted in all kernel methods. For LRF-ELM and LRF-KELM, the random convolutional window and pooling window size are 11×11 and 15×15 respectively. All experiments parameters are selected by cross-validation. Table 1 shows all the comparable accuracies. As the results show, Spectral-LRF-KELM can hardly achieve good classification performance as only considered spectral convolution. It is obvious that LRF-ELM and LRF-KELM outperforms other methods in Fig. 2.

Fig. 3 plots the global accuracy (OA) impact as a results of different convolution and pooling window size. As we can see, larger window size may not achieve better classification performance. From Fig. 3, it can be seen that the best classification performance is achieved when the convolutional window size is 11 and pooling window size is 15 respectively. We also investigate the influence of PCA operations in LRF-KELM by analyzing OA under different PC numbers. The PCA is used twice for dimension reduction. In Fig. 4, we call the first PCA operation as pre PCA and second PCA as post PCA. It is apparent that the OA is climbing with the PC number increasing, which means using more features is better than less features. The differences of classification accuracy

Table 1. CLASSIFICATION ACCURACY (%) FOR THE INDIAN PINES IMAGE ON THE TEST SET

class	Train	Test	KSVM	SVM-CK	KELM	Spectral-LRF-ELM	BF-ELM	LRF-ELM	LRF KELM
Alfalfa	6	48	43.13	61.46	33.33	68.75	93.75	100	**100**
Corn-no till	144	1290	83.43	93.75	79.22	84.19	95.89	94.58	**98.19**
Corn-min till	84	750	75.79	91.39	64.93	64.13	98.27	**97.87**	97.07
Com	24	210	67.52	83.24	53.33	56.19	93.33	98.16	**99.54**
Grass/pasture	50	447	93.24	94.12	92.17	92.39	97.99	90.07	**99.77**
Grass/tree	75	672	95.68	97.22	98.21	95.54	97.17	98.48	**99.70**
Grass/pasture-mowed	3	23	36.96	36.09	30.43	30.43	91.30	**100**	88.00
Hay-windrowed	49	440	99.00	99.11	99.54	98.86	99.77	100	**100**
Oats	2	18	2.78	2.22	55.56	61.11	77.78	100	**100**
Soybeans-no till	97	871	76.76	91.03	69.57	76.46	96.67	**98.75**	98.64
Soybeans-min till	247	2221	87.87	95.16	85.10	88.52	98.69	96.90	**97.62**
Soybeans-clean till	62	552	85.51	91.96	80.80	79.89	91.30	**98.50**	94.37
Wheat	22	190	97.79	97.74	99.46	**99.47**	90.53	96.76	99.46
Woods	130	1164	96.93	98.42	96.22	96.22	99.48	100	**100**
Bidg-grass-tree-drives	38	342	57.37	76.52	64.62	65.20	99.71	98.88	**100**
Stone-stell towers	10	85	87.53	93.76	74.12	83.53	**95.19**	83.13	89.16
Overall Accuracy (%)			74.21	93.37	82.54	84.76	97.29	97.23	**98.29**
Average Accuracy (%)			85.68	81.45	73.54	77.56	94.81	97.00	**97.95**
k			0.758	0.924	0.800	0.825	0.974	0.968	**0.981**



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

(e) (f) (g) (h)

Fig. 2. Thematic maps resulting from classification for the Indian Pines data. (a)Ground truth; (b) KSVM; (c) SVM-CK; (d) KELM; (e) Spectral-LRF-ELM; (f) BF-ELM; (g)LRF-ELM; (h)LRF-KELM.

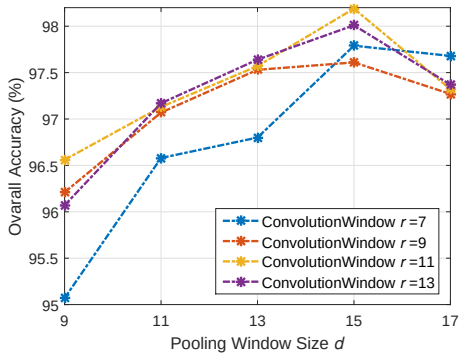


Fig. 3. Influence of convolution window and pooling window size.

is very small when pre PC and post PC are near 26 and 200 receptively.

5. CONCLUSION

In this paper, a method based on LRF-KELM is proposed for HSI classification. In particular, the local spectral-spatial feature of HSI is extracted with random convolution nodes as input layer, which can further improve classification performance by combining KELM. By comparing the LRF-KELM with other classification methods, obvious and numerical merit can be demonstrated in real data experiments.

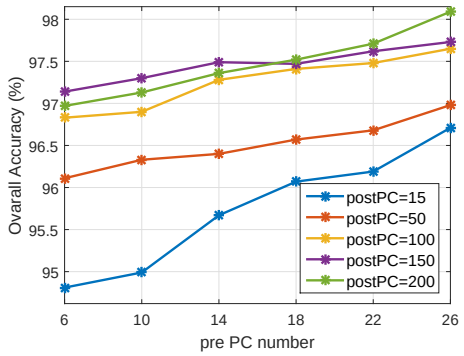


Fig. 4. Influence of two PCA operations in LRF-KELM.

6. REFERENCES

[1] Melgani F, Bruzzone L, "Classification of hyperspectral remote sensing images with support vector machine," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 835-3841, 2004.

[2] Li Jun, J. M. Bioucas-Dias, and A. Plaza, "Hyperspectral Image Segmentation Using a New Bayesian Approach with Active Learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3947-3960, 2011.

[3] Huang Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489-501, 2006.

[4] Pal Mahesh, "Extreme-learning-machine-based land cover classification," *International Journal of Remote Sensing*, vol. 30, no. 14, pp. 3835-3841, 2009.

[5] Pal, Mahesh, Aaron E. Maxwell, and Timothy A. Warner, "Kernel-based extreme learning machine for remote-sensing image classification," *Remote Sensing Letters*, vol.4, no.9, pp. 853-862, 2013.

[6] Bau, Tien C., Subhadip Sarkar, and Glenn Healey, "Hyperspectral region classification using a three-dimensional Gabor filterbank," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 9, pp. 3457-3464, 2010.

[7] Li Wei, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3681-3693, 2015.

[8] Huang Guang-Bin, *et al*, "Local receptive fields based extreme learning machine," *IEEE Computational Intelligence Magazine*, vol. 10, no. 3, pp. 18-29, 2015.

[9] Lv Qi, *et al*, " Hyperspectral image classification via kernel extreme learning machine using local receptive fields," *IEEE International Conference on Image Processing (ICIP)*, pp. 256-260, 2016.

[10] Rodarmel, Craig, and Jie Shan, "Principal component analysis for hyperspectral image classification," *Surveying and Land Information Science*, vol. 62, no. 2, pp. 115-122, 2002.

[11] Camps-Valls, Gustavo, and Lorenzo Bruzzone, "Kernel-based methods for hyperspectral image classifi-cation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351-1362, 2005.

[12] Camps-Valls, Gustavo, *et al*, "Composite kernels for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 1, pp. 93-97, 2006.

[13] Shen Y, Xu J, Li H, *et al*, "ELM-based spectral-spatial classification of hyperspectral images using bilateral filtering information on spectral band-subsets," *IEEE International Geoscience and Remote Sensing Symposium(IGARSS)*, pp. 497-500, 2016.