Spectral-Spatial Classification of Hyperspectral Imagery Based on Stacked Sparse Autoencoder and Random Forest

Chunhui Zhao, Xiaoqing Wan, Genping Zhao, Bing Cui, Wu Liu and Bin Qi

Abstract

It is of great interest in exploiting spectral-spatial information for hyperspectral image (HSI) classification at different spatial resolutions. This paper proposes a new spectral-spatial deep learning-based classification paradigm. First, pixel-based scale transformation and class separability criteria are employed to measure appropriate spatial resolution HSI, and then we integrate the spectral and spatial information (i.e., both implicit and explicit features) together to construct a joint spectral-spatial feature set. Second, as a deep learning architecture, stacked sparse autoencoder provides strong learning performance and is expected to exploit even more abstract and high-level feature representations from both spectral and spatial domains. Specifically, random forest (RF) classifier is first introduced into stacked sparse autoencoder for HSI classification, based on the fact that it provides better tradeoff among generalization performance, prediction accuracy and operation speed compared to other traditional procedures. Experiments on two real HSIs demonstrate that the proposed framework generates competitive performance.

1. Introduction

Hyperspectral imagery has been widely available as a result of advances in remote sensors, and it has high spectral resolution and contains abundant spectral information [Heras et al., 2014; Li et al., 2015]. This characteristic makes it possible to detect and discriminate the subtle differences among land cover classes. The classification of HSI plays an essential role in many applications including both economical and military domains [Plaza et al., 2009; Amato et al., 2013; Mahesh et al., 2015]. Meanwhile, the high dimensionality and complexity of spectral data sets have stimulated the development of several advanced methodologies, such as the well-known support vector machine (SVM) classifier and sparse representation-based classification (SRC). These two kind of classifiers and their derivatives have been widely used for addressing the ill-posed classification problems of high dimensional data space [Melgani and Bruzzone, 2004; Guo et al., 2008; Qian et al., 2012]. In general, these classification methods take full advantage of the spectral information of HSI, whereas the dependencies between adjacent pixels are ignored. Namely, the rich spatial information at neighboring locations may be buried in HSI. To further improve the classification performance, extensive studies have been developed to incorporate spatial information of HSI, based on the assumption that pixels in a local region generally belonging to the same class [Li et al., 2012; Fang et al., 2014].

The above-mentioned methodologies adopt hand-crafted descriptors or transform-based filters to improve classification performance. Nevertheless, the generalization ability may be restricted in certain environments such as insufficient training samples and limited computing units. Motivated by the multiple stages of processing task for human brains, a fast learning algorithm for deep belief nets is proposed [Hinton et al., 2006]. Since then, deep learning model has been demonstrated to be an extremely powerful tool in image processing domain. For instance, object recognition [Chen et al., (2008a)], face verification [Huang et al., 2012], and organ detection [Shin et al., 2013]. Recently, deep learning model has been extended to remote sensing image classification [Chen et al., 2014; Yue et al., 2015].

As a matter of fact, in addition to these well-known classifier models have become the focus of attention in remote sensing domain, it is becoming obvious that the factor of scales (or spatial resolution) plays an increasingly important role in determining the final classification results [Markham and Townshend, 1981]. Woodcock and Strahler (1987) revealed that the local variance of an image changes as the resolution-cell size changes can support in choosing an appropriate image scale. The remote sensing images based on different spatial resolutions significantly affect the statistical characteristic between ground objects [Marceau et al., 1994].

CONTACT Chunhui Zhao, zhaochunhui@hrbeu.edu.cn College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China

© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Further observation also illustrated that the counter effect of spatial resolution on the classification errors associated with with-class variability and boundary effect [Hsieh et al., 2001]. To be brief, spatial resolution is closely related to classification of ground objects, that is, the classification accuracy is the result of a tradeoff of two opposing factors such as boundary effect and with-class variability. Hence, a further research to measure appropriate spatial resolution image is needed in order to enhance the classification performance.

In this paper, a deep learning-based for the spectral-spatial classification paradigm is proposed. It fully considers several constraint factors in HSI classification, including the information derived from the appropriate spatial resolution scene, the classifier models to be used to extract the information, and the spatial structure of pixels in a local region. Firstly, the original remote sensing image is extended to different spatial resolutions by employing pixel-based scale transformation method. Then a spectral angle-based separability criterion is adopted to determine appropriate spatial resolution remote sensing image, where most of ground objects exist the highest average class separability. Moreover, the appropriate scale data and original spectral information are integrated to generate a new feature set. Secondly, considering that the pixels within a local region usually have local similarity. The nearest neighbor domain information of pixels is incorporated to form spectral-spatial vectors. Finally, stacked sparse autoencoder (SSA) is exploited to extract layer-wise more abstract and deep-seated features from original spectral feature set, new feature set, and spectral-spatial vectors. In addition, to further improve the classification performance of HSI, random forest (RF) is first introduced into SSA for hyperspectral data classification. RF exhibits strong generalization performance and high computational efficiency in comparison with bagging, classification and regression tree (CART), and neural network classifier [Ham et al., 2005; Abe et al., 2012]. Last but not least, these obtained high-level features are subsequently fed to RF classifier, and then the classification is determined by a majority vote. To sum up, the proposed spectral-spatial deep learning-based classification framework makes full use of these characteristics: 1) incorporate both spectral and spatial information is a helpful way to enhance the classification performance; 2) as a deep learning architecture, stacked sparse autoencoder provides strong learning performance and is aimed to learn more concise and high-level feature representations than the handcrafted features in describing the underlying data; random forest has shown well generalization performance and high computational efficiency for complex classification problems.

The rest of the paper is organized as follows: Section 2 integrates the spectral and spatial information (i.e., both implicit and explicit features) in order to improve the classification performance, and provides a detailed description of sparse autoencoder (SA), stacked sparse autoencoder and random forest (SSARF), and Welch’s t test; input data sets and quantitative metrics are given in Section 3; the experimental results and performance evaluation provided by different model parameters are demonstrated in Section 4; and, finally, concluding remarks for the present study and future research direction are drawn in Section 5.

### 2. Overview of the methodologies

The general flowchart of the proposed methodology is shown in Figure 1. The first step is the determination of appropriate spatial resolution remote sensing image, and two approaches are adopted to extract the implicit spatial information, including the pixel-based scale transformation method and spectral angle-based separability criterion. Furthermore, the explicit spatial information is extracted by incorporating the nearest neighbor domain information of pixels, considering that the pixels within a local region usually belonging to the same material. The objective of the second step is to generate more abstract and high-level feature representations by using SSA architecture. In the third step, a RF classifier is introduced.

---

**Figure 1.** Block diagram of the proposed methodology.
into SSA in order to determine the probability of being classified into the corresponding land cover class. Finally, we adopt Welch’s t test to verify the statistical significance in classification accuracy improvement of the proposed classification framework.

2.1. The related theoretical background

As briefly mentioned in Introduction, many studies have demonstrated that incorporating both spectral and spatial information in the interpretation of a hyperspectral image is a helpful way to enhance the classification performance. Motivated by the fact that spatial features which can be utilized as discriminant features to supplement the spectral ones [Xia et al., 2015]. The paper proposes a joint spectral-spatial classification framework by integrating the spatial information (i.e., both implicit and explicit spatial information) and spectral information.

2.1.1. Scale transformation

In remote sensing images, spatial resolution is analogous to the scale of observations that is used for describing the detail level of information, that is, it depicts the minimum size of the ground objects that can be separately distinguished or measured [Woodcock and Strahler, 1987]. The classification accuracy of different ground objects was largely influenced by the relative relationship between spatial resolution and target size in environment [Ming et al., 2008; Bai and Wang, 2004].

In this section, the original remote sensing image is up-scaled to different spatial resolutions by using cubic convolution sampling method. Compared with other pixel-based scale transformation methods, cubic convolution method considers the values of the 16 adjacent pixels to calculate intermediate grey-level values between sampled points, that is, it not only efficiently utilizes the elevation values of direct adjacent points, but also considers the influence of the rate of change of elevation values from farther adjacent points [Zhou and Zhang, 2011]. The cubic convolution sampling method can effectively avoid the discontinuity phenomena of ground objects distribution, resulting in the sampling grey values are closer to the original sampling values. Here, assuming that the floating point coordinate \((i + m, j + n)\), target pixel value is denoted as

\[
f(i + m, j + n) = A \cdot B \cdot C
\]

\[A = [K(m + 1) \ K(m) \ K(m - 1) \ K(m - 2)]
\]

\[
B = \begin{bmatrix}
  f(i - 1, j - 1) & f(i - 1, j + 0) & f(i - 1, j + 1) & f(i - 1, j + 2) \\
  f(i + 0, j - 1) & f(i + 0, j + 0) & f(i + 0, j + 1) & f(i + 0, j + 2) \\
  f(i + 1, j - 1) & f(i + 1, j + 0) & f(i + 1, j + 1) & f(i + 1, j + 2) \\
  f(i + 2, j - 1) & f(i + 2, j + 0) & f(i + 2, j + 1) & f(i + 2, j + 2)
\end{bmatrix}
\]

\[C = [K(n + 1); K(n); K(n - 1); K(n - 2)]
\]

In (1), \(m\) and \(n\) respectively represent the floating point numbers, and the interval is \((0, 1)\). The floating point coordinate \(f(i + m, j + n)\) can be obtained through inverse transform the original coordinate \(f(i, j)\). In (5), \(K(x)\) is the approximation value for \(\sin(x \times \pi)/x\).

2.1.2. Class separability criteria

Class separability, being the measure of the similarity between the classes in a feature subspace, and large class separability represents small within-class scattering but large between-class scattering [Tolpekin and Stein A, 2009]. It is broadly studied in literature as an alternative criterion relevant to classification error rate. Essentially, there are several commonly quantitative measures for the class separability such as divergence, scatter-matrix-based measure, B-distance, and J-M-distance [Wang, 2008]. Basically, separability measurement criteria need to meet several properties such as simple computation and be appropriate for the classification task, unfortunately, most of the existing separability criteria are difficult to match these demands [Ge and Mo, 2006]. The scatter-matrix-based measure is often restricted to positively evaluate the class separability in case of the data presents a non-Gaussian structure. The
probability distribution function based criteria such as divergence and B-distance, these measures are computationally intensive due to the computation of covariance matrix and corresponding inversion is involved in these terms. In addition, a further reason for explaining they are not applicable for hyperspectral classification is that the number of instances per class is generally less than the number of features, resulting in the covariance matrix is singular and is unable to compute the corresponding inverse.

To obtain a scene with the maximum homogeneity within region and maximum heterogeneity between regions, the spectral angle-based separability criterion is adopted to measure appropriate scale remote sensing image. In other words, the average within-class spectral angle and average between-class spectral angle are used to calculate class separability for each two-class problem. In comparison to above mentioned commonly used class separability measures, the spectral angle-based separability criterion can efficiently avoid the constraint problem, that is, the insufficient training samples are not efficient to calculate covariance matrix under the condition of high dimension bands involved [Chen et al., (2008b)].

Assume that two-class ground objects are $X$ and $Y$, respectively. It corresponds to band subset $k$. The average spectral angle of within-class $S_W$, average spectral angle between-class $S_B$, and the total average spectral angle $S_T$ are respectively represented by Equations (6)-(8)

$$S_W = \frac{1}{2N_X} \sum_{i=1}^{N_X} \theta_{SAM}^k(x_i, \bar{x}) + \frac{1}{2N_Y} \sum_{i=1}^{N_Y} \theta_{SAM}^k(y_i, \bar{y})$$  \hspace{1cm} (6)

$$S_B = \frac{1}{N_XN_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \theta_{SAM}^k(x_i, y_j)$$  \hspace{1cm} (7)

$$S_T = S_W + S_B$$  \hspace{1cm} (8)

$$J(\cdot) = \frac{S_B}{S_B + S_W}$$  \hspace{1cm} (9)

where $N_X$ and $N_Y$ denote the number of samples in class $X$ and class $Y$, respectively. $\theta_{SAM}^k$ represents the subangle of band subset $k$. In (9), criterion function $J(\cdot)$ is adopted to simultaneously minimize $S_W$ and maximize $S_B$. In the ideal case, the greater $J(\cdot)$ can be interpreted as better statistical separability of the two classes will be.

The implicit spatial information from appropriate scale remote sensing image is extracted by performing the pixel-based scale transformation method and spectral angle-based separability criterion, to thereby generate a new feature set by integrating the implicit spatial information and original spectral information. In addition, to enhance the classification performance, we further incorporate the explicit spatial information of new feature set. It should be noted that a large neighborhood region will bring about large input dimension for HSI classification. Hence, we incorporate nearest neighbor domain information of pixels by performing a post-processing procedure for each category label, based on the assumption that the nearest neighbor domain information is relevant with original pixels. Finally, the original spectral feature set, new feature set, and the joint spectral-spatial feature vectors are prepared for the SSARF architecture, shown within the green dashed-line box in Figure. 4.

2.2. Deep learning, SA, and SSARF

2.2.1. Deep learning

Deep learning is an emerging approach in machine learning field, which is widely used to learn multiple levels of abstract representation of data by training a multilayer neural network. In comparison to most traditional machine learning methods (such as linear SVM, SRC, and softmax regression), which are attributed to single-layer classifiers by using shallow-structured learning frameworks, while deep learning (DL) references the multiple stages of processing for human brains that uses unsupervised and supervised patterns to hierarchically learn feature representations in deep networks for classification task. There are two main reasons for employing DL architecture in classification. First of all, DL is a hierarchical learning mechanism which extracts in-depth features from just the pixel intensities alone by stacking unsupervised networks on top of another, that is, the main advantage of model is to discover the distributed feature representation of data. Furthermore, the obtained features are considered useful for identifying different ground objects by cascading a superior classifier. The second reason is that it allows to automatically extract the features from the pixel intensities in a recursive manner and doesn’t need design the handcrafted features by trial and error, and hence the model is expected to be much more robustness and the pretreatment time can be largely reduced [Guo et al., 2016].

For the past few years, deep learning has received an increasing interest in the machine learning research communities since the first learning algorithm for deep belief nets is proposed [Hinton, et al., 2006]. Since then, the deep learning algorithms have shown the competitive performance to learn useful feature representations directly from data for a wide variety of domains [Lenz et al., 2013; Yue et al., 2015]. As previously described, it is necessary to automatically learn the feature representations from the data in place of utilizing handcrafted features. To achieve this, stacked sparse autoencoder (SSA) is adopted to adaptively extract the abstract
feature representations from the pixel intensities. Essentially, as a hierarchical learning architecture, SSA is constructed by stacking multi-layer sparse autoencoders and each layer is trained in turn with the greedy layerwise algorithm (see Figure 2). In the following sections, the autoencoder as a basic unsupervised artificial neural network is first described. Then, we introduce the shallow sparse autoencoder by imposing a sparsity constraint and a weight decay term on the network so as to discover interesting structures in data. Furthermore, we focus on learning the high-level feature representations by using stacked sparse autoencoder. Finally, a random forest (RF) classifier is introduced into the SSA for hyperspectral image classification. The proposed classification model is expected to be useful to discover the distributed feature representations of data and achieve strong generalization performance, and then enhance the prediction accuracy and calculation stability.

2.2.2. Sparse Autoencoder (SA)
A shallow sparse autoencoder presents a special type of neural networks consisting of one input layer, one hidden layer and one reconstruction layer, which can be utilized to learn the abstract features from the labeled and unlabeled datasets in an unsupervised learning principle [Hinton and Salakhutdinov, 2015]. In other words, to learn the intrinsic features within the dataset in a mapping weight and bias vectors pattern, SA tries to learn approximation to identity function so as to reconstruction vector at the decoding layer is going to be similar to input vector at the input layer.

In general, the input layer of an autoencoder which maps an input \( x \in \mathbb{R}^D \) to the corresponding representation \( z \in \mathbb{R}^S \), and the hidden layer \( z \) can be regarded as a abstract feature representation of the input vector (see Eq.10).

\[
z = f(W_z x + b_z)
\]  

(10)

where \( W_z \in \mathbb{R}^{S \times D}, \ b_z \in \mathbb{R}^{S \times 1} \) \((D\) represents the dimension of input data, and \( S \) denotes the number of neurons in the hidden layer) represent the weights and biases of the input layer to hidden layer, respectively. To generate a nonlinear mapping, the logistic sigmoid function \( f(x) = (1 + \exp(-x))^{-1} \) is employed in both the encoder and decoder. Moreover, the hidden representation \( z \) is used to reconstruct an approximation \( y \) of the input \( x \) via a decoder from output layer (see Eq.11).

\[
y = f(W_y z + b_y)
\]  

(11)

where \( W_y \in \mathbb{R}^{D \times S}, \ b_y \in \mathbb{R}^{D \times 1} \) represent the weights and biases of the hidden layer to output layer, respectively. For rendering the parameterizations identical, we restrain \( W_y = W_z^T \). Basically, the optimal features are extracted by minimizing the error \( \cos t(\cdot) \) between raw input and reconstruction, as shown in Equation (12).

\[
\arg\min_{W_y, \ b_y} \cos t(y, x)
\]  

(12)

For \( M \) input samples, the reconstruction error \( \cos t(y, x) \) can be re-expressed by \( C(Y, X) = \frac{1}{M} \sum_{i=1}^{M} \left( \frac{1}{2} \| y_i - x_i \|_2^2 \right) \), where \( Y \) and \( X \) represents the reconstruction and input over the \( M \) training samples, respectively. Specifically, to discover interesting structure in raw input whether hidden units \( S \) is smaller or larger than the dimension of input data \( D \), the overall energy function \( I_{cos} \) in the sparse autoencoder is to minimize the reconstruction error with a sparsity constraint and a weight decay terms (see Eq.13).

\[
I_{cos} = \frac{1}{M} \sum_{i=1}^{M} \left( \frac{1}{2} \| y_i - x_i \|_2^2 \right) + \frac{\lambda}{2} \sum_{l} \sum_{i} \sum_{j} (W_{ij}^{(l)})^2 + \eta \sum_{j=1}^{S} KL(r \| \bar{r}_j)
\]  

(13)

where the first term on the right side of Equation is an average sum-of-squares error term which represents the discrepancy between input \( x \) and reconstruction \( y \) [Tao et al., 2015]. The second term denotes the weight decay term, which is employed to reduce the autoencoder from overfitting by controlling the amplitude of the weights, where \( \lambda \) is a weight decay parameter, and \( W_{ij}^{(l)} \) denotes the connection between the \( i^{th} \) unit in layer \( l-1 \) and the \( j^{th} \) unit in layer \( l \) [Zhang et al., 2015]. The third term denotes a sparsity penalty term, where \( \eta \) controls the weight of the term, and \( KL(r \| \bar{r}_j) \) is a Kullback-Leibler divergence for measuring how different between rand \( \bar{r}_j \) (see Eq. 14).

\[
KL(r \| \bar{r}_j) = \sum_{j=1}^{S} \left( r \log \frac{r}{\bar{r}_j} + (1-r) \log \frac{1-r}{1-\bar{r}_j} \right)
\]  

(14)
where \( r \) is the parameter of desired sparsity, and \( \tilde{r}_i = \frac{1}{M} \sum_{j=1}^{M} |y_j(x_i)| \) is the average activation of hidden unit \( j \) over the training data \( x_i \), that is, the sparsity is minimized since \( \tilde{r}_i \) is close to \( r \) for each hidden unit.

The minimization process of the objective function \( J_{\text{cost}} \) can be implemented by using stochastic gradient descent and backpropagation algorithm [Liu and Nocedal, 1989; Rumelhart et al., 1986]. It is worth noting that reconstruction layer is removed and the learned features are encapsulated in hidden layer after learning the network. And then the obtained features are utilized as input vectors of the next layer so as to extract high-level features.

### 2.2.3. Stacked sparse autoencoder and random forest (SSARF)

#### A. Hierarchal learning architecture SSA

By using SA, the low-level features can be learned from the raw data in a hierarchal manner. However, the obtained low-level features are not adequate because of the large appearance variations of the hyperspectral data. Inspired by the multiple stages of processing for human brains, we adopt stacked sparse autoencoder (SSA) to progressively extract more abstract and higher-order features of data based on the low-level features, and the architecture is believed to has the ability of generating promising performance than those shallower classifiers [Varga et al., 2015]. In this section, a typical SSA is constructed by stacking the input and hidden layers of SAs layer by layer (see Figure 2), and it can be trained by adopting greedy methods for each additional layer [Bengio et al., 2007; Ng, 2011]. For simplicity, the decoder parts of each shallow SA would not be displayed in Figure 2. Similar to the learning scheme of SA, the purpose of training a SSA is to achieve the optimized weight and bias values by minimizing the error between raw input and its reconstruction, and then the outputs of each layer except the last one are wired to the inputs of next layer. The specific implementation process of SSA architecture is given as follows.

Set \( W_z \) and \( W_y \) represent the input-to-hidden and the hidden-to-output weight matrices, respectively. Here \( b_z \) and \( b_y \) represent the biases of hidden and output layers, respectively. Firstly, the low-level features \( z^{(1)}_i \) is learned by training the first SA over the input \( x_i \), where \( z^{(1)}_i = f(W_z^{(1)} x_i + b_z^{(1)}) \). Secondly, the outputs of the first SA are regard as the inputs of next layer, which is encoded by the second SA in order to obtain the high-level feature representations \( z^{(2)}_i \), i.e., \( z^{(2)}_i = f(W_z^{(2)} z^{(1)}_i + b_z^{(2)}) \). By parity of reasoning, we employ the recursive encoding procedure to extract the \( l^{th} \) layer feature representations \( z^{(l)}_i = f(W_z^{(l)} z^{(l-1)}_i + b_z^{(l)}) \), \( l > 1 \). To sum up, the first layer of SSA tends to learn first-order feature representations in raw input, and the second layer tends to learn second-order feature representations on the basis of first-order features, then by parity of reasoning, the higher layers tend to learn even higher-level feature representations [Varga et al., 2015]. Particularly, after achieving encoding procedure for each layer network, the decoding procedure of SSA is performed to reconstruct the input vectors of each additional layer. In the section, the low-level features \( y^{(l-1)}_i \) can be reconstructed by the high-level feature representations \( y^{(l)}_i \), i.e., \( y^{(l-1)}_i = f(W_y^{(l)} y^{(l)}_i + b_y^{(l)}) \), \( l > 1 \). Furthermore, the raw input can be reconstructed by the low-level reconstructions \( y^{(1)}_i \), i.e., \( y_i = f(W_y^{(1)} y^{(1)}_i + b_y^{(1)}) \).

By performing above-mentioned steps, the more abstract and high-level feature representations are achieved by layer-wise pre-training in an unsupervised manner. Once the whole architecture of SSA is trained, the output features of the highest layer are considered to be more beneficial to the classification task. To further make the learned features discriminative, the last decoder layer is removed and a random forest classifier is added on the top of the last hidden layer of the SSA. Then we further fine-tune the whole network before performing classification. Crucially, the number of nodes in the classification output layer equivalents to the number of category labels, rather than the number of the neurons in the hidden layers.

#### B. The output layer: random forest (RF) classifier

In recent times, various ensemble methods (such as bagging, boosting and random subspace) have been developed to predict land cover for un-sampled map units and have been proven to be beneficial to retrieve important information from the scene [Ramirez et al., 2009]. Bagging, as a widely used ensemble method, is based on bootstrapping. But the boosting method is based on sample re-weighting technique. In addition, the random subspace ensemble, is an ensemble construction technique by integrating diverse component classifiers. Indeed, Random forest (RF) is considered to be one of the most popular random subspace ensemble methods, and it is originally proposed by Leo Breiman [Breiman, 2001]. The RF is a decision tree ensemble method based on bagging and random subspace, which can be utilized to address such problems as high-dimen-sional data and high feature-to-instance ratio [Kuncheva et al., 2010]. In the past few years, due to the important advantages of RF in terms of generalization performance and computational efficiency, and has become a superior classification tool and is often used in fields ranging from remote sensing to medical science [Ham et al., 2005; Speiser et al., 2014]. In this period, studies have demonstrated that RF classifier generates the better performance compared to other procedures such as Bagging, Boosting, neural network classifier, and classification and regression tree (CART)
classifiers [Ham et al., 2005; Abe et al., 2012]. Furthermore, the related literature has experimentally compared the performances of the well known SVM and RF classifiers, and result showed that the classification accuracy of RF classifier is comparable with that of SVM classifier and RF reduces training and testing costs significantly over a regular SVM [Bosch, et al., 2007]. Hence, we introduce a RF classifier into the SSA architecture so as to predict outcomes, based on the important advantages such as well generalization performance, high prediction accuracy and fast operation speed. The main implementation process behind the RF machine learning algorithm is as follows.

RF is a soft classifier of decision tree based ensemble methods, as shown in (Figure. 3). The model is a majority vote mechanism of decision tree predictors in which each tree is formed using resampling technique with replacement. Consequently, different subsets from the original training set are adopted to form each tree (in-bag set). Meanwhile, the remaining is put down the tree to construct a test classification (out-of-bag set). Furthermore, the best splits are selected among random subset of the predictor variables, in which case a terminal node occurs. To classify the out-of-bag dataset in RF classifier, the vector assignment of class label of an unknown instance is determined by a majority vote. Basically, the RF algorithm is based on the Gini index minimum principle (see Eq. 15), and the Gini index is described by

\[
\text{Gini}(s) = \sum_{i=1}^{K} p_{n_i} (1 - p_{n_i})
\]  

(15)

where \(K\) is the number of classes, and \(p_{n_i}\) is the probability of being classified into the corresponding land cover class \(n_i\) at node \(s\), as defined by Equation (16)

\[
p_{n_i} = \frac{m_{n_i}}{m}
\]  

(16)

where \(m_{n_i}\) is the number of trees belonging to class \(n_i\), and \(m\) is the total number of classification trees.

C. Implementation

As can be seen from (Figure. 4), firstly, we adopt the pixel-based scale transformation method and class separability criterion to measure the appropriate scale remote sensing image. Secondly, the nearest neighbor domain information of pixels (both original data and appropriate scale data) is incorporated, based on the assumption that neighbor pixels are more similar than pixels far away. The red dot represents a random pixel,
and four blue dots represent its nearest neighbor domain information. Finally, the spectral-spatial vectors are fed to SSARF model in order to exploit meaningful and high-level features from both spectral and spatial domains, meanwhile improving generalization performance and computational efficiency in HSI classification. The algorithm flowchart is drawn up in the following.

After executing the spectral-spatial classification from all trained classifiers, we apply Welch’s t-test to further demonstrate whether the proposed methodology shows the significant improvement compared with traditional classifiers in terms of classification accuracy. A brief description of the Welch’s t test is given in Section 2.3.

2.3. Welch’s t test

Comparison of locations, or central tendency, of two independent treatments are widely exist in the present computer simulation research [Fagerland and Sandvik, 2009]. The Welch’s t test (also called unequal variances t-test) is the most common method in statistics, and is a two-sample location test. At present, it has been utilized as a representative metric for discriminating polarized measurements based on two populations are normally distributed with unequal variances. Basically, Welch’s t test is a modification of the Student’s t test, which has been demonstrated with the assist of Student’s t-test and is more robust when there is a possibility of unequal variances and unequal sample sizes between the two treatments [Shrestha et al., 2013]. Moreover, the choice of these tests decides what can be achieved from the independent populations, based on the different null hypotheses are constructed to test. In other words, these tests are often regarded as “unpaired” or “independent samples” t-tests, have been widely used when the statistical units from the two treatments being compared are non-overlapping.

Notice that there are two main reasons why we choose Welch’s t test rather than the traditional t test (which assumes equal variances) to verify the statistical significance in classification accuracy improvement. First, the Welch’s t-test performs as well as, or more robust than, the Mann-Whitney-Wilcoxon test (i.e., a non-parametrical method) and the Student’s t-test with the Type I and Type II error rates whenever the underlying distributions are normal [Ruxton, 2006]. Second, many literatures have demonstrated that even when unequal variances are combined with non-normality distributions, the Mann-Whitney-Wilcoxon test and the traditional t test make the type I error rates strongly deviate from nominal level. Nevertheless, Welch’s t test would still be available for analyzing two-sample comparisons and shown robustness to non-normality in the underlying populations [Keselman et al., 2004; Ruxton, 2006]. The implementation of Welch’s t test is as follows.

We assume that the classification accuracy $C$ of the proposed method conforms to the normal distribution with mean $\mu$ and variance $\sigma^2$. The classification accuracy $C_0$ based on benchmark framework also conforms to the normal distribution with mean $\mu_0$ and variance $\sigma_0^2$. Thus, we could have $C - C_0 \sim N(\mu - \mu_0, \sigma^2/m + \sigma_0^2/m_0)$, where $m$ and $m_0$ denote the number of land-cover classes for the proposed method and benchmark classification frameworks, respectively. The hypothesis is expressed as

$$K : \mu > \mu_0 \quad K_0 : \mu \leq \mu_0$$ (17)

In this section, we apply one-tailed test to execute statistical inference, based on the reason that a priori hypothesis is expected by (Equation (17)), where one-tail test is allowed to reject priori hypothesis even when the difference between these populations is relatively small. The test statistic of two-sample t test is given as

$$T = \frac{\bar{u} - \mu_0}{\sqrt{\frac{s^2}{m} + \frac{s_0^2}{m_0}}}$$ (18)

where $\bar{s}^2$ and $s_0^2$ are replaced with their sample variance $S^2$ and $S_0^2$, respectively. Therefore, the test statistic of Welch’s t test can be rewritten as

Figure 4. A deep learning-based framework for spectral-spatial classification. The framework is used for feature learning and classification. The architecture within the black dotted-line box represents the random forest ensemble model.
\[ T = \frac{(u - u_0)}{\sqrt{(S^2/m + S_0^2/m_0)}} \]  

(19)

To execute the hypothesys (see Eq. 17), the distribution of the above test statistic is approximated as an ordinary Student’s t distribution. Here, Welch-Satterthwaite is employed to approximate the degree of freedom \( \nu \) associated with this variance estimate.

\[ \nu \approx \frac{(S^2/m + S_0^2/m_0)^2}{\left( \frac{S^4}{(m^2(m-1))} + \frac{S_0^4}{(m_0^2(m_0-1))} \right)} \]  

(20)

The probability to reject \( K_0 \) conditioned on valid \( K_0 \) is given as

\[ H(\text{reject } K_0|\text{valid } K_0) = H(T \geq t_a(v)) \]

\[ = H\left( \frac{u - u_0}{\sqrt{S^2/m + S_0^2/m_0}} \geq t_a(v) \right) \]

(21)

In (21), the significant level \( \alpha \) is set to be 0.1, and it is used to determine the quantile \( t_a(v) \). The situation of accepting hypothesis \( K_0 \) is

\[ H(T \geq t_a(v)) \leq \alpha \]

\[ = H\left( \frac{u - u_0}{\sqrt{S^2/m + S_0^2/m_0}} \geq t_a(v) \right) \leq \alpha \]  

(22)

In (22), we require that the probability error cannot exceed significant level \( \alpha \), where the critical value \( t_a(v) \) is obtained by looking up the Student’s t distribution table [Pearson and Hartley, 1967]. Moreover, the statistical values of Welch’s t-test of the proposed classification algorithm and contrastive algorithms are calculated by Equation (19). Once significant level \( \alpha \) and the degree of freedom \( \nu \) have been obtained, these statistics can be used to test the alternative hypothesis that one of the population means is greater than the other.

3. Input data sets and quantitative metrics

3.1. Experimental data sets

The experiments are conducted on two commonly-used hyperspectral data sets, namely Indian Pines scene and Kennedy Space Center (KSC). Indian Pines scene was captured by Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor in Northwestern Indiana. This scene has 220 data channels across the spectral range from 0.2 to 2.4 \( \mu \)m, with a spatial resolution of 20 m per pixel [Fang et al., 2014]. In the experiments, 20 water absorption and noisy bands (no. 104-108, 150-163, and 220) are removed and the remaining 200 bands are used for the analysis [Mirzapour and Ghassemian, 2015]. Indian Pines scene contains 16 different land-cover classes, and a pseudo color image corresponds to different land cover classes are visually shown in (Figure. 5 (a)-(b)). It is worthy noting that a few factors resulting in the land-cover classes in the scene may be difficult to be effectively identified. On the one hand, among the 16 classes, “corn-notill”, “corn-mintill”, and “corn” are the same species of corn with different types. In addition, “grass-pasture”, “grass-trees”, and “grass-pasture-mowed” are regarded as similar land-cover classes and have similar spectral properties. “Soybean-notill”, “soybean-mintill”, and “soybean-clean” are the same species of soybean with different types. Furthermore, the deviation degree of the number of samples maintains large in some classes, such as the “soybean-mintill” and “oats”. On the other hand, Indian Pines image is not a large spatial structure (i.e., 144×144 pixels) and the adjacent interval of different land-cover classes is relative small. This may leads to the presence of highly mixed pixels in boundary region that complicates the classification problem, especially for some areas where the number of samples of land-cover classes is small. Thus, in order to make the experimental analysis more significant, seven classes have a very small number of elements are removed and the remaining nine classes ranging in size from 497 to 2468 pixels are considered. In this section, we randomly select 10% of each class as the training samples and the rests for testing, which is presented in (Table.1).

The KSC data set was acquired by the NASA AVIRIS over the Kennedy Space Center in Florida. The KSC scene consists of 224 bands with center wavelengths from 400 to 2500 nm. The image size is 512 \( \times \) 614, with a spatial resolution of 18 m. Low SNR bands are removed and the remainder 176 bands are used for experiments. A false color image and the reference classes are described in (Figure.6 (a)-(b)). For classification purposes, 13 land-cover classes with 5211 pixels that occur in this environment were described, and the labeled samples vary in size from 105 to 927 pixels. We randomly choose 10%
of the labeled samples for training and use the rest samples for testing. The number of training and test samples for each class is listed in (Table. 2).

3.2. Quantitative metrics

In this section, we conduct many groups of experiments to quantitatively evaluate the feasibility and effectiveness of proposed classification framework. The kappa coefficient ($\kappa$) and overall accuracy (OA) are used to evaluate the quality of classification results. OA is the percentage of correctly classified pixels in the whole scene. The kappa is a robust metric for categorical data and is commonly used in remote sensing applications. 

![Figure 5](image) For Indian Pines image: (a) Pseudocolor image, (b) Ground truth, and (c) Information classes. Classification maps obtained by (d) OS-SVM, (e) NF-SVM, (f) OS-SRC, (g) NF-SRC, (h) OS-SSARF, (i) NF-SSARF, and (j) SSF-SSARF.

![Figure 6](image) For KSC image: (a) Pseudocolor image, (b) Ground truth, and (c) Information classes. Classification maps obtained by (d) OS-SVM, (e) NF-SVM, (f) OS-SRC, (g) NF-SRC, (h) OS-SSARF, (i) NF-SSARF, and (j) SSF-SSARF.

**Table 2.** Information classes and training-test sets for KSC data set.

| Class                  | Train | Test  |
|------------------------|-------|-------|
| C1- Scrub              | 76    | 685   |
| C2- Willow swamp       | 24    | 219   |
| C3- CP hammock         | 25    | 231   |
| C4- CP/Oak             | 25    | 227   |
| C5- Slash pine         | 16    | 145   |
| C6- Oak/Broadleaf      | 22    | 207   |
| C7- Hardwood swamp     | 10    | 95    |
| C8- Gaminoid marsh     | 43    | 388   |
| C9- Spartina marsh     | 52    | 468   |
| C10- Cattail marsh     | 40    | 364   |
| C11- Salt marsh        | 41    | 378   |
| C12- Mud flats         | 50    | 453   |
| C13- Water             | 92    | 835   |
| **Total**              | **516** | **4695** |

---

Table adapted from PDF content.
measure of degree of agreement which integrates diagonal and off-diagonal entries of confusion matrix. Note that, to simplify the description, in the process of executing SVM, SRC and SSARF models. The classification performances based on original spectral feature set are defined as OS-SVM, OS-SRC and OS-SSARF, respectively; the classification performances based on the new feature set are denoted as NF-SVM, NF-SRC and NF-SSARF, respectively. Finally, under the condition of performing the SSARF model, the classification performance based on the spectral-spatial feature set is represented as SSF-SSARF. Specially, all experiments were executed using MATLAB on a personal computer with i5-4210M CPU@2.6GHz Core, 4GB of RAM.

4. Experimental results and discussion

4.1. Experiments on Indian Pines dataset

The first experiment is performed on the Indian Pines data. The appropriate scale image with 40m spatial resolution is utilized in this scenario, and the specific selection criteria for the role will be explained in the next section. In order to evaluate the proposed classification framework SSF-SSARF, a comparison with SVM and SRC is adopted as the contrastive algorithms. In the case of SVM, we adopt radial basis function (RBF) kernel function and one-against-one strategy for M-class classification. The RBF kernel parameter and regularization parameter are selected by 10-fold cross validation. The sparsity-based classification algorithm is based on the assumption that the spectral signatures of pixels in the same class lie in a low-dimensional subspace and then a test sample can be sparsely represented by a linear combination of small number of dedicated atoms in the training dictionary. To re-construct the sparse representation, we employ the orthogonal matching pursuit (OMP) algorithm to deal with the sparsity-constrained optimization problem so as to determine the class label of the test sample. In terms of OS-SSARF, NF-SSARF, and SSF-SSARF models, a number of trials with changing the configuration are conducted to select the suitable parameter values. The detailed process will be illustrated in the Section 4.3 and 4.4.

The classification quantitative evaluation results generated by different classification frameworks are shown in (Table. 3). It can be seen that NF-SVM, NF-SRC and NF-SSARF achieve better classification results compared with OS-SVM, OS-SRC and OS-SSARF. They cause increase in the overall classification accuracy of 1.18%, 1.02% and 1.19%, respectively. As expected, SSF-SSARF obtains the highest classification performance among all the classification frameworks. The effectiveness of classification performance can be further validated by visually observing the classification maps (see Figure. 5). The unambiguous results are obtained for spectral classification maps which have a very noisy appearance. Nevertheless, the NF-SVM, NF-SRC, NF-SSARF and SSF-SSARF obviously reduce the noise interference.

| Class          | OS-SVM | NF-SVM | OS-SRC | NF-SRC | OS-SSARF | NF-SSARF | SSF-SSARF |
|----------------|--------|--------|--------|--------|----------|----------|-----------|
| Corn-notill    | 84.30  | 84.14  | 68.55  | 69.25  | 84.43    | 85.92    | 89.22     |
| Corn-mintill   | 76.26  | 80.92  | 59.46  | 62.65  | 77.36    | 78.52    | 82.43     |
| Grass-pasture  | 95.58  | 96.87  | 91.63  | 89.89  | 93.97    | 94.59    | 94.81     |
| Grass-trees    | 97.97  | 96.61  | 97.69  | 97.49  | 96.14    | 96.09    | 97.69     |
| Hay-windrowed  | 99.17  | 98.75  | 99.32  | 99.45  | 99.77    | 98.57    | 99.08     |
| Soybean-notill | 75.80  | 83.19  | 72.88  | 75.72  | 81.08    | 82.86    | 80.96     |
| Soybean-mintill| 87.83  | 87.87  | 76.62  | 78.14  | 86.61    | 90.50    | 90.67     |
| Soybean-clean  | 80.95  | 82.35  | 57.77  | 59.28  | 80.47    | 84.17    | 88.61     |
| Woods          | 98.38  | 98.32  | 97.99  | 97.54  | 99.06    | 98.83    | 96.95     |
| OA (%)         | 87.75  | 88.93  | 78.85  | 79.87  | 86.57    | 89.76    | 90.67     |
| \( \kappa \)   | 0.8555 | 0.8698 | 0.7518 | 0.7637 | 0.8655   | 0.8797   | 0.8904    |

4.2. Experiments on KSC dataset

The second experiment is executed on the KSC data. The appropriate scale image with 36m spatial resolution is utilized in this scenario. Implementation procedures are the same to Indian Pines data. As can be seen from (Table. 4), it is found that NF-SVM, NF-SRC and NF-SSARF outperform OS-SVM, OS-SRC and OS-SSARF, respectively. The improvement for NF-SVM, NF-SRC and NF-SSARF in OA are 1.1%, 0.90% and 1.63%, respectively. It is also demonstrated that the SSF-SSARF tends to more robust than NF-SSARF, which obtains the highest classification accuracies for most of classes. The classification maps of various classification frameworks are shown in (Figure. 6). As can be observed, the proposed SSF-SSARF further reduces the noise and generates the best classification map among all the classification frameworks.

4.3. Effect of class separability

In remote sensing image classification, appropriate scale remote sensing image provides the best compromise between detail of changes detected and the size of resultant data volume. To verify this, the Indiana Pines scene with a spatial resolution of 20 m per pixel is progressively resampled to coarser
spatial resolutions by using pixel-based scale transformation method. Specially, the resampling process is performed by merging pixels within $s \times s$ grid into a single window or a larger grid with cubic convolution resampling method, where $s$ represents the scale factor. For Indiana Pines scene, to obtain different spatial resolutions (i.e., 40m, 60m, 80m, 100m and 120m), $s$ is set to be 2, 3, 4, 5, and 6, respectively. Moreover, the above-mentioned resampling method is also applies to the KSC scene. Under the condition of different spatial resolutions, appropriate scale remote sensing image can be determined by using class separability criteria. Similar strategy has also demonstrated that the optimal resolution remote sensing image could be determined by using statistical separability, and the finer spatial resolution does not necessarily lead to high separability [Bai and Wang, 2004]. In general, for N land-cover classes, $N \times (N-1)/2$ transformed divergences for each pair of classes that must be constructed to evaluate how statistical separability changes with different spatial resolutions [Bai and Wang, 2004]. In other words, $9 \times (9-1)/2$ and $13 \times (13-1)/2$ class sepearabilities that must be constructed for each pair of classes since nine classes have a large number of elements are considered in Indian Pines scene and 13 land-cover classes are described in the KSC scene. To simplicity, we take the mean value of class separabilities between each class and other classes as a rough measurement standard, which is denoted as average class separability (ACS) (see Figure. 7). Note that the average class separabilities of all information classes are further averaged to be used as the overall separability ‘M’ of the input data set.

As can be seen from (Figure. 7), the maximum ACS of each class exists in different spatial resolution images. For Indiana Pines scene, the maximum ACS of some classes exists in original spatial resolution scene, such as Grass-pasture, Woods. However, the maximum ACS of most of classes is generated from 40 m spatial resolution scene. Furthermore, the ACS of each class generally shows downward trend with spatial resolution continues to increase. For KSC scene, most of classes acquire highest ACS in 36 m spatial resolution image. The experiment results illustrate that the finer spatial resolution does not necessarily generate strong the statistical separability.

### Table 4. Individual class accuracy values (in percent), overall accuracy (OA) and kappa coefficient ($\kappa$) of different classification frameworks for KSC data set. The highest result of each row is highlighted in boldface.

| Class            | OS-SVM | NF-SVM | OS-SRC | NF-SRC | OS-SSARF | NF-SSARF | SSF-SSARF |
|------------------|--------|--------|--------|--------|----------|----------|-----------|
| Scrub            | 93.87  | 95.47  | 94.89  | 95.00  | 93.82    | 95.39    | 97.68     |
| Willow swamp     | 97.26  | 91.51  | 83.56  | 85.87  | 88.28    | 92.13    | 93.17     |
| CP hammock       | 89.61  | 96.40  | 85.28  | 88.66  | 91.63    | 92.54    | 96.68     |
| CP/Oak           | 80.18  | 73.24  | 48.02  | 50.18  | 66.52    | 77.23    | 71.30     |
| Slash pine       | 66.90  | 63.19  | 60.00  | 63.19  | 64.60    | 72.16    | 71.98     |
| Oak/Broadleaf    | 65.22  | 57.98  | 43.48  | 44.19  | 71.82    | 73.02    | 80.15     |
| Hardwood-swamp   | 71.58  | 75.83  | 61.05  | 82.20  | 78.25    | 84.72    | 89.17     |
| Graminoid-marsh  | 87.89  | 90.35  | 86.60  | 84.54  | 92.18    | 93.77    | 92.92     |
| Spartina marsh   | 94.87  | 96.93  | 91.45  | 93.50  | 97.08    | 97.67    | 99.91     |
| Cattail marsh    | 92.86  | 90.35  | 95.60  | 90.95  | 96.15    | 95.10    | 97.48     |
| Salt marsh       | 85.71  | 98.51  | 93.39  | 96.83  | 96.56    | 96.73    | 98.72     |
| Mud flats        | 91.17  | 95.25  | 81.68  | 80.35  | 95.88    | 97.13    | 94.64     |
| Water            | 98.32  | 97.79  | 97.72  | 98.37  | 100.00   | 100.00   | 100.00    |
| OA (%)           | 90.01  | 91.11  | 85.73  | 86.63  | 91.84    | 93.47    | 94.43     |
| $\kappa$         | 0.8888 | 0.9008 | 0.8409 | 0.8509 | 0.9091   | 0.9273   | 0.9378    |

![Figure 7](image_url) The mean value of class separabilities for each class with other classes and the mean of average class separabilities of all information classes, are depicted versus different spatial resolutions (a) Indiana Pines data set (b) KSC data set.
4.4. Model parameters estimation

4.4.1. Effect of different percentages of training samples

In order to avoid biasing conclusions caused by the number of training samples, a set of experiments is executed to observe the effects of different training samples for the whole classification frameworks. The training sample set is provided by randomly choosing different percentages (from 2% to 12% for each class) of samples, and the rest are considered as test samples. In each experiment, to ensure that the different classification frameworks are effectively evaluated in a fair and stable way, and the accuracy values are averaged over ten repetitions. The results are illustrated in Figure 8. It can be observed that for both images, the accuracy gradually improves with the increase of the number of training samples for each classification framework. As expected, the SSF-SSARF is obviously outperforms other classification frameworks under the condition of keeping the same training samples.

4.4.2. Effect of model depth

When process and analyze Indian Pines and KSC data sets by using SSF-SSARF model, different model depths have a great effect on the classification performance. Here, model depth refers to the number of hidden layers in deep learning-based architecture, and it plays an important role in classification task due to it can determine the quality of feature representations from raw data. In general, when the model depth is set to be 1, the SSA architecture tends to extract first-order feature representations in raw input. If the model depth is set to be 2, the SSA can learn two-level feature representations. In other words, the higher layers tend to extract more abstract and even higher-order feature representations. Thus, to verify this viewpoint, a set of experiments is performed to evaluate how the model depth of the feature representations influences the classification performance. Specially, experiments are performed on the basis of keeping prerequisite for consistency, that is, input data set, the number of hidden layer neurons, and iterations are fixed. The experimental results are shown in Table 5.

As can be seen from Table 5, for two commonly-used hyperspectral data sets, the overall classification accuracy gradually improves with the increase of model depths, and it roughly exhibits downward trend with the model depth continues to increase. In general, the execution time (i.e., both training time and testing time) rapidly grows with the increase of model depths. In the ideal condition, we prefer to employ an available model that generates the highest classification accuracy and spends the least amount of execution time. In the case of compromising the classification accuracy and the execution time for Indiana Pines and Kennedy Space Center scenes, the model depths (or the number of hidden layers in the SSF-SSARF architecture) are set to be 2 and 3, respectively, where the classifier model obtains the highest classification accuracy.

![Figure 8. Classification performance of each classification framework with different percentages of training samples (a) Indian Pines data set (b) KSC data set.](image-url)

![Table 5. Comparison of overall accuracy for different model depths.](table-url)
4.4.3. Evaluation of execution time
The execution time of deep learning-based architecture consists of two parts, such as the training time and the testing time. As can be seen from (Figure 9), 10% of the labeled samples in two commonly-used hyperspectral data sets for training, and the training time refers to the time consumption of training each hidden layer, classification layer, and fine-tune the whole network. In other words, the training time, that is, the time needed to complete the maximum pre-training epochs and fine-tuning epochs in the process of unsupervised pre-training and supervised fine-tuning. Broadly speaking, Figure 9 is used to evaluate how the training time changes with the change of the model parameters (such as the number of hidden neurons and iterations). It can be observed that the training time gradually grows with the increase of the number of hidden layer neurons and iterations. Here, network size refers to the size of hidden units (or the number of hidden neurons). For example, the KSC scene consists of 13 land-cover classes and 176 spectral channels, that is, input size is set to be 176. If each hidden layer size in SSA architecture is set to 60, and also a classification layer (i.e., random forest classifier) is employed on the top of the SSA architecture. Thus, the deep learning-based architecture are constructed as 176-60...60-13. In addition, in the process of training the SSARF architecture and predicting the class labels, the number of iterations refers to the maximum pre-training epochs or fine-tuning epochs.

![Figure 9](image-url) Comparison of training time under the condition of different hidden sizes (i.e., the number of hidden neurons) and iterations (a) Indian Pines data set (b) KSC data set.

4.5. Evaluation of statistical significance
To verify the statistical significance in classification accuracy improvement of the SSF-SSARF

| Model depth |
|-------------|
| 1 | 2 | 3 | 4 | 5 |
| Testing time (s) | AVIRIS | 0.41 | 0.45 | 0.47 | 0.50 | 0.54 |
|              | KSC | 0.25 | 0.28 | 0.32 | 0.34 | 0.37 |

Table 6 shows the comparison of testing time for different model depths. In table 6, 90% of the unlabeled samples for testing, that is, 8416 pixels in Indian Pines and 4695 pixels in Kennedy Space Center data set are used for classifying the class labels, where the testing time is compared after performing unsupervised pre-training and supervised fine-tuning. It can be seen that an advantage of the proposed SSF-SSARF architecture is that it shows super-fast on testing, and the time is less than a second.

Specially, the testing time is much less than the training time, and the major reason is that the training stage includes unsupervised pre-training and supervised fine-tuning. However, the testing stage doesn’t need retraining the deep learning-based model and only provides the corresponding class labels for the unlabeled samples. In the case of compromising the classification accuracy and execution time for Indiana Pines data set. The pre-training iterations are set to be 950 for each layer, whereas the fine-tuning iterations are set to be 2000, and the hidden size is set to be 100. For KSC scene, the pre-training iterations are set to be 800 for each layer, meantime the fine-tuning iterations are set to be 2000, and the hidden size is set to be 60.

Table 6. Comparison of testing time for different model depths.
classification framework, Welch’s t test is performed in this experiment. Figure 10 shows the difference values between statistical values and critical values with different classification algorithms, where the training samples vary in percentage from 2% to 12%. If the probability error (i.e., this ratio between the number of difference values on the horizon and the number of all difference values) is less than or equal to the significant level, where hypothesis $K_0$ is tenable. Otherwise, we will refuse to the hypothesis $K_0$. It can be seen from Figure 10 that most of statistical values of Welch’s t test fall in the reject region, and hence we will accept another hypothesis $K$ that the proposed SSF-SSARF has an obvious improvement compared with the contrastive frameworks.

5. Conclusions

In this paper, a new spectral-spatial mechanism is exploited on the stacked sparse autoencoder-based model for hyperspectral image classification. The original spectral data and appropriate scale data is incorporated as a new feature set to increase the class separability among different ground object classes. Experimental results show that the new feature set helps to increase the classification accuracy compared with original spectral data for each classifier model. In addition, SSARF model acquires the higher classification accuracy and stronger robust performance than other classifier techniques like SVM and SRC. Meanwhile, the cascaded spectral-spatial vectors which integrate nearest neighbor information of pixels obtain the best classification performance in SSF-SSARF model. Finally, Welch’s t-test further verifies the proposed SSF-SSARF framework outperforms the state-of-the-art related frameworks in terms of classification accuracies. In the process of measuring appropriate scale remote sensing image, apart from class separability criterion, shape characteristics also have a certain influence on the expression of typical ground objects. In the latter work, we will consider how to effectively utilize shape characteristics to further improve the classification performance.

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China (Nos. 61405041 and 61571145), the China Postdoctoral Science Foundation (Grant No. 2014M551221), the Key Program of Heilongjiang Natural Science Foundation (No. ZD201216), the Program Excellent Academic Leaders of Harbin (No. RC2013XK009003), and the Fundamental Research Funds for the Central Universities (No. HEUCF1608).

References

Heras D.B., Argüello F., Quesada-Barriuso P. (2014) - Exploring ELM-based spatial-spectral classification of hyperspectral images. International Journal of Remote Sensing, 35(2): 401–423. doi: http://dx.doi.org/10.1080/01431161.2013.869633.
Li W., Chen C., Su H., Du Q. (2015) - Local Binary Patterns and Extreme Learning Machine for Hyperspectral Imagery Classification. IEEE Transactions on Geoscience and Remote Sensing, 53(7): 3681–3693. doi: http://dx.doi.org/10.1109/TGRS.2014.2381602.
Plaza A., Benediktsson J.A., Boardman J.W., Brazile J., Bruzzone L., Camps-Valls G., Chanussot J., Fauvel M., Gamba P., Gualtieri A., Marconcini M., Tilton J.C., Trianni G. (2009) - Recent advances in techniques for hyperspectral image processing. Remote Sensing of Environment, 113(1): 110–122. doi: http://dx.doi.org/10.1016/j.rse.2007.07.028.
Amato U., Antoniadis A., Carfora M.F., Colandrea P., Cuomo V., Franzese M., Pignatti S., Serio C. (2013) - Statistical Classification for Assessing PRISMA Hyperspectral Potential for Agricultural Land Use. IEEE
Breiman L. (2001) - *Random Forests*. Machine Learning, 45 (1):5–32.

Kuncheva L.I., Rodriguez J.J., Plumpton C.O., Linden D.E., Johnston S.J. (2010) - *Random subspace ensembles for FMRI classification*. Medical Imaging IEEE Transactions on, 29(2):531–42. doi: http://dx.doi.org/10.1109/TMI.2009.2037756.

Speiser J.L., Durkalski V.L., Lee W.M. (2014) - *Random forest classification of etiologies for an orphan disease*. Statistics in Medicine, 34(5): 887–899. doi: http://dx.doi.org/10.1002/sim.6351.

Bosch A, Zisserman A, Muoz X. (2007) - *Image Classification using Random Forests and Ferns*. IEEE International Conference on Computer Vision. IEEE Computer Society, 1–8.

Fagerland M.W., Sandvik L. (2009) - *Performance of five two-sample location tests for skewed distributions with unequal variances*. Contemporary Clinical Trials, 30 (5):490–496. doi: http://dx.doi.org/10.1016/j.cct.2009.06.007.

Shrestha S., Giakos G., Farrahi T., Livanos G., Zervakis M. (2013) - *Statistical analysis on polarimetric study of lung cancer cells*. IEEE Imaging Systems and Techniques, 34:121–126.

Ruxton G.D. (2006) - *The unequal variance t-test is an underused alternative to Student’s t-test and the Mann–Whitney U test*. Behavioral Ecology, 17(4):688–690. doi: http://dx.doi.org/10.1093/beheco/ark016

Keselman H.J., Othman A.R., Wilcox R.R., Fradette K. (2004) - *The new and improved two-sample t test*. Psychological Science, 15(1):47–51.

Mirzapour F, Ghassemian H. (2015) - *Improving hyperspectral image classification by combining spectral, texture, and shape features*. International Journal of Remote Sensing, 36(4): 1070–1096. doi: http://dx.doi.org/10.1080/01431161.2015.1007251.

Pearson E.S, Hartley H.O. (1967) - *Biometrika Tables for Statisticians*. Volume I. (Third Edition). Journal of the American Statistical Association, 62 (319).