We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

5,000
Open access books available

125,000
International authors and editors

140M
Downloads

154
Countries delivered to

TOP 1%
Most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
Chapter

Hyperspectral Image Classification

Rajesh Gogineni and Ashvini Chaturvedi

Abstract

Hyperspectral image (HSI) classification is a phenomenal mechanism to analyze diversified land cover in remotely sensed hyperspectral images. In the field of remote sensing, HSI classification has been an established research topic, and herein, the inherent primary challenges are (i) curse of dimensionality and (ii) insufficient samples pool during training. Given a set of observations with known class labels, the basic goal of hyperspectral image classification is to assign a class label to each pixel. This chapter discusses the recent progress in the classification of HS images in the aspects of Kernel-based methods, supervised and unsupervised classifiers, classification based on sparse representation, and spectral-spatial classification. Further, the classification methods based on machine learning and the future directions are discussed.

Keywords: hyperspectral imaging, classification, supervised and unsupervised classification, machine learning

1. Introduction

The technological progression in optical sensors over the last few decades provides enormous amount of information in terms of attaining requisite spatial, spectral and temporal resolutions. Especially, the generous spectral information comprises of hyperspectral images (HSIs) establishes new application domains and poses new technological challenges in data analysis [1]. With the available high spectral resolution, subtle objects and materials can be extracted by hyperspectral imaging sensors with very narrow diagnostic spectral bands for the variety of purposes such as detection, urban planning [2], agriculture [3], identification, surveillance [4], and quantification [5, 6]. HSIs allow the characterization of objects of interest (e.g., land cover classes) with unprecedented accuracy, and keep inventories up to date. Improvements in spectral resolution have called for advances in signal processing and exploitation algorithms.

Hyperspectral image is a 3D data cube, which contains two-dimensional spatial information (image feature) and one-dimensional spectral information (spectral-bands). Especially, the spectral bands occupy very fine wavelengths, while the image features such as Land cover features and shape features disclose the disparity and association among adjacent pixels from different directions at a confident wavelength.

In the remote sensing community, the term classification is used to denote the process that assigns individual pixels to a set of classes. The output of the classification step is known as the classification map. With respect to the availability of training samples, classification approaches can be split into two categories, i.e.,
supervised and unsupervised classifiers. Supervised approaches classify input data for each class using a set of representative samples known as training samples. Hyperspectral (HS) image classification always suffers from varieties of artifacts, such as high dimensionality, limited or unbalanced training samples [7], spectral variability, and mixing pixels. The Hughes phenomenon is a common problem in the supervised classification process [8]. The power of classification increases with the increase of available training samples. The limited availability of training samples decreases the classification performance with the increase of feature dimension. This effect is famously termed as “Hughes phenomenon” [9]. It is well known that increasing data dimensionality and high redundancy between features might cause problems during data analysis. There are many significant challenges that need to be addressed when performing hyperspectral image classification. Primarily, supervised classification faces challenge about the imbalance between high dimensionality and incomplete accessibility of training samples or the presence of mixed pixels in the data [10]. Further, it is desirable to integrate the essential spatial as well as spectral information so as to combine the complementary features that stem from source images [11]. A considerable amount of literature has been published with regard to overcoming these challenges, and performing hyperspectral image classification effectively.

Hyperspectral image classification could attract scientific community which aims at assigning a pixel (or a spectrum) to one of a certain set of predefined classes. Maximum likelihood (ML) methods, neural networks architectures [12], support vector machine (SVM) [13], Bayesian approach [14] as well as kernel methods [15] are the prominent methods which have been investigated in recent years for the identification or classification of hyperspectral data.

Based on the usage of training sample, image classification task is categorized as supervised, unsupervised and semi-supervised hyperspectral image classification.

2. Unsupervised classification

The paramount challenge for HSI classification is the curse of dimensionality which is also termed as Hughes phenomenon. To confront with this difficulty, feature extraction methods are used to reduce the dimensionality by selecting the prominent features. In unsupervised methods, the algorithm or method automatically groups pixels with similar spectral characteristics (means, standard deviations, etc.) into unique clusters according to some statistically determined criteria. Further, unsupervised classification methods do not require any prior knowledge to train the data. The familiar unsupervised methods are principal component analysis (PCA) [16] and independent component analysis (ICA) [17].

2.1 Principal component analysis

It is the most widely used technique for dimensionality reduction. In comparative sense, appreciable reduction in the number of variables is possible while retaining most of the information contained by the original dataset. The substantial correlation between the hyperspectral bands is the basis for PCA. The analysis attempts to eliminate the correlation between the bands and further determines the optimum linear combination of the original bands accounting for the variation of pixel values in an image [18].

The mathematical principle of PCA relies upon the eigen value decomposition of covariance matrix of HSI bands. The pixels of hyperspectral data are arranged as a
vector having its size same as the number of bands. \( X_i = [x_1, x_2, \ldots, x_N]^T \), where \( N \) is the number of HS bands. The mean of all the pixel vectors is calculated as:

\[
m = \frac{1}{M} \sum_{i=1}^{M} [x_1 \ x_2 \ \ldots \ x_N]^T
\]

where \( M = p \times q \) is the number of pixel vectors for a HS image of "p" rows and "q" columns. The covariance matrix is determined as:

\[
C = \frac{1}{M} \sum_{i=1}^{M} (X_i - m)(X_i - m)^T
\]

(2)

The covariance matrix can also be written as:

\[
C = ADA^T
\]

(3)

\( D \) is the diagonal matrix composed of eigen values \( \{\lambda_1, \ldots, \lambda_N\} \) of \( C \) and \( A \) is the orthogonal matrix with the corresponding eigen vectors (each of size \( N \)) as columns. The linear transformation \( y_i = A^TX_i \), \( i = 1, 2, \ldots, M \), is adapted to achieve the modified pixel vectors which are the PCA transformed bands of original images. The first \( K \) rows of the matrix \( A^T \) are selected such that, the rows are the eigen vectors corresponding to the eigen values arranged in a descending order. The selected \( K \) rows are multiplied with the pixel vector \( X_i \) to yield the PCA bands composed of most of the information contained in the HS bands.

In hyperspectral data, most of the elements are covered by the sensors with high spectral resolution which cannot be well described by the second order characteristics. Hence, PCA is not an effective tool for HS image classification since it deals with only second-order statistics.

2.2 Independent component analysis (ICA)

Independent component analysis successfully executes the independence of the components with higher-order statistics, and is relatively more suitable to encounter high dimensionality of HS images. ICA is an attractive tool for dimensionality reduction, feature extraction, blind source separation, etc., as well as to preserve the information which cannot be retrieved using second order statistics [19, 20].

Let us consider a mixture of random variables \( x_1, x_2, \ldots, x_N \), where each \( x_i \in \mathbb{R}^d \). These random variables are defined as a linear combination of another random variables \( p_1, p_2, \ldots, p_N \), where each \( p_i \in \mathbb{R}^n \). In such scenario, the mixing model can be mathematically written as,

\[
X = AP
\]

(4)

where \( X = [x_1, x_2, \ldots, x_N] \) is the observed vector, \( P = [p_1, p_2, \ldots, p_N] \) is the unknown source, \( A \) is the mixing matrix, “n” denotes the number of unknown sources and “d” represents the number of observations made. In order to find the independent components, the unmixing matrix \( W \) is to be estimated (inverse of \( A \)). The independent components are obtained using Eq. (5).

\[
\text{ICA}(X) = P = A^{-1}X = WX
\]

(5)
If \( X \in \mathbb{R}^{d \times N} \) is considered as the hyperspectral image,

\[
P_{n \times N} = W_{n \times d}X_{d \times N}
\]  

where \( N \) is the number of pixels in each band, \( d \) represents the number of spectral bands and \( n \) gives the number of sources or materials present in the image. The estimation of the ICA model is conceivable, only if the following presumptions and limitations are fulfilled: (i) Sources should be statistically independent (ii) Independent components should possess non-Gaussian distribution (iii) Matrix \( A \) should be a square and full rank matrix.

3. Supervised classification

The supervised classification takes the advantage of rich spectral information and has explored many applications including urban development [21], the monitoring of land changes [22], target detection [23], and resource management [24]. In supervised classification only labeled data is used to train the classifier. A large number of supervised classification methods have been discussed in the literature, some of the prominent methods are maximum likelihood (ML), nearest neighbor classifier, decision trees, random forest, support vector machines (SVMs), etc.

Figure 1 shows the conventional steps of supervised classification of HSIs.

3.1 ML classifier

The ML classifier assumes that the statistics for each class in each band are normally distributed and estimates the probability that a given pixel belongs to a certain specific class [25]. Unless a probability threshold is selected, all pixels are classified. Each pixel is assigned to a particular class that manifests the maximum probability. If the estimated maximum probability is smaller than a threshold, the pixel remains unclassified. The following discriminant functions for each pixel in the image are implemented in ML classification.

![Flowchart of HSI supervised classification.](image-url)
\[
g_i(x) = \ln p(w_i) - \frac{1}{2} \ln |\sigma_i| - \frac{1}{2} (x - m_i)^T \sigma_i^{-1} (x - m_i) \tag{7}
\]

where \(i = \text{class}\); \(x = \text{n-dimensional data (where n represents the number of bands)}\); \(p(w_i) = \text{probability that class } w_i \text{ occurs in the image and is assumed the same for all classes}\); \(|\sigma_i| = \text{determinant of the covariance matrix of the data in class } w_i\); \(\sigma_i^{-1} = \text{its inverse matrix}\); and \(m_i = \text{mean vector}\).

Implementation of the ML classification involves the estimation of class mean vectors and covariance matrices using training pattern chosen from known examples of each particular class [26]. It usually acquires higher classification accuracy compared to other traditional classification approaches. It assumes that each band is normally distributed and the chosen training samples are comprised of exhaustively defined set of classes. For hyperspectral data with tens of hundreds of spectral bands, discrimination of land cover classes is not an easy task, whereas, the classification accuracy of ML classifier is based on the accurate selection of the training samples. Thus, for the hyperspectral imagery with poorly represented labeled training samples, it is preferable to adapt an alternative to the standard multiclass classifier.

### 3.2 k-nearest-neighbor (kNN) classifier

kNN is one of the widely used simplest classifier, and has been applied for HSI classification [27, 28].

kNN method operates on majority voting rule, presumes that all the neighbors make equal contributions to the classification of the testing point. Another important feature of kNN classifier is Euclidian is used as distance metric, which assumes the data is homogeneous.

Let \(X = [x_1, \ldots, x_N]\) be the N-point training data, with \(d\) as the dimension of each point. \(X_i = [x_{i1}, \ldots, x_{ik}]\) be the \(k\) nearest neighbors of \(x_i\). The testing data (\(N_t\) points) is denoted as \(X_t\) with \(x_0\) is a random testing point. The \(k\) nearest neighbors from the testing data with labels \([l_1, l_2, \ldots, l_k]\) is indicated as \(X_0 = [x_{01}, \ldots, x_{0k}]\). Let assume that \(\Omega = [\Omega_1, \ldots, \Omega_C]\) are the \("C"\) classes in the data.

The kNN classifier finds the \(k\) nearest neighbors of a testing point in the training data and assigns the testing point to the most frequently occurring class of its \(k\) neighbors. The classification of \(x_0\) by majority voting rule is exercised using the following expression:

\[
j^* = \arg \max_{j=1, \ldots, C} \sum_{i=1}^{k} \delta(l_i, j) \tag{8}
\]

where \(\delta\) is the Kronecker delta.

A distance metric learned from the given training data is used to enhance the accuracy of kNN classifier.

\[
dis(x_i, x_j) = \|T(x_i - x_j)\|^2 \tag{9}
\]

\(T\) denotes a linear transformation.

The decision rule of kNN can be modified by assigning different weights to the neighbors. Further, the testing point is assigned to the class for which the sum of weights chosen for the neighbors is largest.

\[
j^* = \arg \max_{j=1, \ldots, C} \sum_{i=1}^{k} w_i \delta(l_i, j) \tag{10}
\]
It is also referred as decision rule for weighted $k$NN (W$k$NN), where $w_i$ is the weight of $x_{0i}$.

3.3 Spectral angle mapper (SAM)

SAM is a supervised classification technique for HSIC [29]. SAM classifier admits very quick classification using the spectral angle information of HSI data.

The reference spectra are usually determined from the field measurements or from the image data, is used to measure the spectral angle. The spectral angle is a n-dimensional vector between image and reference spectra. Smaller the angles between two spectrums, higher the similarity and vice versa. The classification approach using SAM is described in Figure 2.

This technique is comparatively insensitive to illumination and albedo effects when reflectance data is used for analysis. The spectral angle can be calculated as follows:

$$\theta = \cos^{-1}\left(\frac{\sum_{i=1}^{N} T_i R_i}{\sqrt{\sum_{i=1}^{N} T_i^2} \sqrt{\sum_{i=1}^{N} R_i^2}}\right)$$  \hspace{1cm} (11)$$

![Figure 2. SAM classification approach.](image)

---

*Processing and Analysis of Hyperspectral Data*
3.4 Support vector machine (SVM)

SVM is typically a linear classifier associative with kernel functions and optimization theory and is prominent for HSI classification [13, 30, 31]. SVM outperforms the conventional supervised classification methods particularly in prevailing conditions like increased number of spectral bands and the limited availability of training samples [32–34].

3.4.1 Linear SVM: Linearly separable case

Let \( x_i \in \mathbb{R}^d, (i = 1, 2 \ldots N) \) be the set of training vectors, and a target \( y_i \in \{-1, +1\} \) is corresponding to each vector \( x_i \). The problem is treated as a binary classification and the two classes are linearly separable. Hence, at least one hyperplane must exist to separate the two classes without errors. The discriminant function associated with hyperplane can be defined as:

\[
 f(x) = wx + b \tag{12}
\]

where \( w \in \mathbb{R}^d \) is a vector normal to hyperplane, \( b \in \mathbb{R} \) is a bias. \( w \) and \( b \) must satisfy the following condition to estimate such a hyperplane,

\[
 y_i(w \cdot x_i + b) > 0, \quad \text{for } i = 1, 2 \ldots N \tag{13}
\]

The optimal hyperplane can be estimated by solving the following convex problem.

\[
 \min \frac{1}{2} ||w||^2 \quad \text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1, \quad \text{for } i = 1, 2 \ldots N \tag{14}
\]

3.4.2 Linearly nonseparable case

For practical data classification problem, the linearly separable condition may not be true in different conditions. To solve the classification problem of nonseparable data, hyperplane separation has been generalized. A cost function is formulated comprising two conditions: margin maximization (as in the case of linearly separable data) and error minimization (to penalize the wrongly classified samples).

\[
 \psi(w, \xi) = \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} \xi_i \tag{15}
\]

Where, \( \xi_i \) are slack variables derived to account for the nonseparability of data and \( C \) is a regularization parameter. The larger the \( C \) value, the higher the penalty associated with misclassified sample.

The minimization of the cost function defined in Eq. (15) is subject to the following conditions:

\[
 y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1, 2 \ldots N. \tag{16}
\]

\[
 \xi_i \geq 0, \quad i = 1, 2 \ldots N. \tag{17}
\]

For nonseparable data, two types of support vectors coexist: (1) margin support vectors that lie on the hyperplane margin and (2) nonmargin support vectors that fall on the “wrong” side of this margin [13].
3.4.3 Nonlinear SVM – kernel method

The effective discriminant function to solve the nonlinear classification problem can be expressed as:

\[ f(x) = \sum_{i \in S} a_i y_i K(x_i, x) + b \]  

(18)

A common example of kernel type that fulfills Mercer’s condition is the Gaussian radial basis function:

\[ K(x_i, x) = \exp \left( -\gamma \|x_i - x\|^2 \right) \]  

(19)

where, \( \gamma \) is a parameter that is inversely proportional to width of the Gaussian kernel. The more details about kernel functions for this case can be referred in [35].

4. Random forest classifier

A random forest (RF) is a group of tree-based classifiers where each tree is trained with a bootstrapped set of training data. The data to be classified is applied as an input to each tree in the forest. The classification given by each tree is known as a “vote” for that class. In the classification, the forest chooses the class having the most votes (over all the trees in the forest). In RF classification a split is determined by searching across a random subset of variables at each node [36, 37].

The Random forest classifier (RFC) features two main characteristics: relatively high accuracy and the speed of processing. However, the correlation/independence of trees can affect the accuracy of final land cover map. The primitive components of Random Forest are explored as:

4.1 CART-like trees

Classification and regression tree (CART), a binary tree in which splits are resolved by the variables obtained from the strong change in impurity or minimum impurity \( \bar{i}(t) \),

\[ \bar{i}(t) = \sum_{i \neq j} \hat{P}(x_i | t) \hat{P}(x_j | t) \]  

(20)

where \( \hat{P}(x_i | t) \) is the estimated probability of sample \( x_i \in \) class i. The definite classification takes place during training process. Either the impurity is zero or all the splits result in only one node then the growth of the tree terminates.

4.2 Binary hierarchy classifier (BHC)

In contrary to CART, the split on each node in BHC is based on classes. The optimal split at each node is based on class separability and further the splits are pure.

Let us consider a single meta-class case, which split into two into 2 meta-classes and so on, until the true classes are realized in the leaves, while simultaneously computing the Fisher discriminant and projection.

Let \( \mu_j \) and \( \sigma_j \in \{ y, \beta \} \) are the estimated mean vector and co-variance matrix of the meta class \( w_j \), then the data projected using \( w \):
The inverse of class covariance matrix $W$ is given by

$$w = W^{-1}(\mu_\alpha - \mu_\beta)$$  \hspace{1cm} (21)$$

The discriminant $T(W)$ can be maximized as:

$$T(w) = \frac{w^TBw}{w^TWw}$$  \hspace{1cm} (22)$$

where $B$ is the covariance matrix between classes.

$$B = (\mu_\alpha - \mu_\beta)(\mu_\alpha - \mu_\beta)^T.$$  \hspace{1cm} (23)$$

Like the CART trees, the BHC trees can be combined as a forest (RF-BHC) to realize an ensemble of classifiers, where the best splits on classes are performed on a subset of the features in the data to diversify individual trees and/or to stabilize the $W$.

5. Spatial-spectral classification

The pixel-wise classification methods incur some difficulties: Discriminating the classes is very difficult due to less interclass spectral variability. If interclass variability is high, it is very hard to determine a given class. The pixel-wise classification capability can be enhanced by the exploration of additional information called spatial dependency. The classification performance can be improved by incorporating spatial information into HSIC. This rationale motivates the study of spatial-spectral classification methodologies [38]. The spatial dependency system for spectral-spatial-based classification is depicted in Figure 3. The spatial dependency (primary information for spatial-spectral classification techniques) is carried by two identities called pixel and associated label. The correlation among spatially related pixels is spatial dependency, hence spatially related pixels are termed as neighboring pixels. The spatial dependency is associated with (i) Pixel dependency indicates the correlation of neighboring pixels and (ii) Label dependency indicates the correlation of labels of neighboring pixels. Distinct approaches of spatial-spectral classification are as follows [39]:

i. Structural filtering: The spatial information from a region of the hyperspectral data is extracted by evaluating the metrics like mean and standard deviation of neighboring pixels over a window. The relevant methods include spectral-spatial wavelet features [40], Gabor features [41], Wiener filtering [42], etc.

ii. Morphological profile (MP): mathematical morphology (MM) intent to investigate spatial relationships between pixels using a set of known shape and size which is called the structuring element (SE). Dilation and erosion are the two elemental MM operations used for nonlinear image processing. The concept of extracting the information regarding contrast and size of the structures present in an image is termed as granulometry. The morphological profile (MP) of size $n$ has been defined as the composition of a granulometry of size $n$ built with opening by reconstruction and a (anti)granulometry of size $n$ built with closing by reconstruction [43].
From a single panchromatic image, the MP results in a \((2n + 1)\)-band image. However, for hyperspectral images the direct construction of the MP is not straightforward, because of the lack of ordering relation between vectors. In order to overcome this shortcoming, several approaches have been considered [44].

i. Random field: random field-based methods have been studied broadly for HSI classification. Markov random fields (MRFs) and conditional random fields (CRFs) are two major variants of RF-based classification methods. CRF methods adapt conditional probability for labeling the data and attain favorable performance by utilizing the optimal spatial information; whereas, MRF-based techniques achieve substantial reduction in computational complexity by estimating class parameters independently from field parameters. The basic formulation of random fields as follows:

\[
\text{MP}^{(n)}(I) = \begin{bmatrix} \phi^{(n)}_1(I), \ldots, \phi^{(1)}_n(I), I, \gamma^{(1)}_1(I), \ldots, \gamma^{(n)}_n(I) \end{bmatrix}
\]  

(25)

Let \( S = \{1, \ldots, n\} \) denote a set of integers indexing the \( n \) pixels of a hyperspectral image. A conditional probability \( P(y|x) \) (a posteriori) is defined with \( x = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^{d 	imes n} \) denotes \( d \)-dimensional feature vectors composes a hyperspectral image and \( y = \{y_1, y_2, \ldots, y_n\} \) is an image of labels. The a posteriori probability can be expressed as:

\[
p(y|x) = \frac{1}{Z(\omega, x)} \exp \left( \sum_{i \in S} \log p(y_i|x, \omega) + \mu \sum_{(i, j) \in C} \delta(y_i - y_j) \right)
\]

(26)

The normalizing factor \( Z(\omega, x) \), also known as partition function is defined as:
\[ Z(\omega, x) = \sum_{\gamma} \exp \left( \sum_{i \in S} \log p(y_i | x_i, \omega) + \mu \sum_{(i, j) \in C} \delta(y_i - y_j) \right) \]  

(27)

where, \( p(y_i | x_i, \omega) \) is the class probability given by the learning parameter \( \omega \).

\( \mu \) = parameter controlling the degree of smoothness on the image of labels.

\( \delta(y) \) = unit impulse function and \( C \) is a set of cliques.

The CRFs not only avoids label bias problem but also its conditional nature motivates the relaxation of independence assumptions. Recently, Distributed Random Forest (DRF) have gained interest for HSIC [45] owing to its inherent merit.

The salient features of DRF are (1) the relaxation of conditional independence of the observed data, (2) the exploitation of probabilistic discriminative models instead of the generative MRFs, and (3) the simultaneous estimation of all DRF parameters from the training data.

6. Sparse-representation (SR)-based classification

The role of SR theory has become prevalent in almost all the image processing applications. The SR theory presumes that the training samples can be represented as a linear combination of smallest possible number of atoms (columns) of an over-complete dictionary.

The test sample \( \mathbf{x}_i \) can be represented as \( \mathbf{x}_i = \mathbf{D} \alpha + \epsilon \), where, \( \mathbf{D} \in \mathbb{R}^{n \times k} \) is a dictionary with \( n \) samples of \( k \) dimensions and the sparse coefficients vector \( \alpha \) can be determined by solving the following optimization problem.

\[
\hat{\alpha} = \arg \min_{\alpha} \|\alpha\|_0 \quad \text{s.t.} \quad \|\mathbf{x}_i - \mathbf{D} \alpha\|_2 \leq \epsilon
\]  

(28)

The term \( \|\cdot\|_0 \) is \( l_0 \) norm that counts the number of nonzero entries. The optimization problem in Eq. (28) can be solved with greedy pursuit algorithms [46], in which the \( l_0 \) norm is replaced with the \( l_1 \) norm.

For HSIC, the Eq. (28) can be replaced as:

\[
\min_{\alpha} \frac{1}{2} \|\mathbf{x}_i - \mathbf{D} \alpha\|_2^2 + \tau \|\alpha\|_1, \quad \alpha \geq 0
\]  

(29)

where, the parameter \( \tau \) is a Lagrange multiplier that balances the tradeoff between the reconstruction error and the sparse solution: \( \tau \to 0 \) when \( \epsilon \to 0 \).

In order to incorporate the spatial information a spatial weight is added and the modified SR model for HSIC is formulated as:

\[
\min_{\alpha} \frac{1}{2} \|\mathbf{x}_i - \mathbf{D} \alpha\|_2^2 + \tau \|\mathbf{W} \alpha\|_1, \quad \alpha \geq 0
\]  

(30)

The choice of a spatial weight matrix \( \mathbf{W} \), yields different classification strategies for HSIs namely neighboring pixels [47], neighboring filtering [38], histogram-based [47], spatial information based on super pixels [48], etc.

The class labels can be implied on the basis of the following formulation:

\[
\text{class}(\mathbf{x}_i) = \arg \min_{j \in \{1, \ldots, c\}} \|\mathbf{x}_i - \mathbf{D} \alpha_j\|_2.
\]  

(31)

A sparsity-based algorithm to improve the classification performance is proposed in [49]. The principle depends on the sparse representation of a hyperspectral...
pixel by a linear combination of a few training samples from a structured dictionary. The sparse vector is recovered by solving a sparsity-constrained optimization problem, and it can directly determine the class label of the test sample. Zhang et al. [50] proposed a nonlocal weighted joint sparse representation (NLW-JSRC) to further improve the classification accuracy. The method enforced a weight matrix on the pixels of a patch in order to discard the invalid pixels whose class was different from that of the central pixel. A few of the recent investigations [51–53] approved that a compact and discriminative dictionary learned from the training samples can significantly reduce the computational complexity.

6.1 Segmentation-based methodologies

The segmentation process is performed after spectral-based classification in some of HSIC techniques. The extraction and classification of homogeneous objects is presented in [54] is the first classifier that used spatial postprocessing. The comprehensive survey of other methodologies of this category is presented in [43].

7. Deep learning (DL)

Deep learning involves a class of models which try to hierarchically learn deep features of input data with very deep neural networks, typically deeper than three layers. The network is first layer-wise initialized via unsupervised training and subsequently, tuned in a supervised manner. In this scheme, high level features are learned from low level ones, whereas, the proper features can be formulated for pattern classification towards the end. Deep models can potentially lead to progressively more abstract and complex features at higher layers, and more abstract features are generally invariant to the most local changes experienced by the input data.

7.1 Deep learning for HSI classification

The DL theory presents a dynamic way for unsupervised feature learning using very large raw image dataset. Unlike the traditional classification techniques, DL-based techniques can represent and organize multiple levels of information to express complex relationships between data.

Deep Learning (DL) is a sort of more complex architecture simulating human brains, based on neural networks begins to apply hyperspectral image classification [55]. The deep learning models for HSIC usually consists of three layers, to extract the more complex characteristics layer by layer. (i) Input data (ii) Deep layer construction (iii) Classification [56]. The notable methodologies include deep belief network (DBN) [57], stacked auto encoder (SAE) [58], and convolutional neural network (CNN) [59].

Deep belief networks (DBNs) [60] are an important development in DL research and train one layer at a time in an unsupervised manner by restricted Boltzmann machines (RBMs) [61]. The DBNs admit unsupervised pretraining over unlabeled samples at first and then a supervised fine-tuning over labeled samples. Since the pretrained DBN captures the useful information from the unlabeled samples, the fine-tuning with the pretrained DBN performs well over small number of labeled samples [57, 62]. The simple structure of DBN is presented in Figure 4.

The conventional training of DBN incur two problems; The first is coadaptation of latent factors [63, 64]. This activity is described as several latent factors tend to behave very similarly. This phenomenon implies that the model parameters
corresponding to the latent factors might be very similar. These similar latent factors make most of the computations to be performed redundantly and also decrease DBN’s description ability. The second is the set of many “dead” (never responding) or “potential over-tolerant” (always responding) latent factors (neurons) in the DBN learned with the usual sparsity promoting priors [65]. The “dead” or “potential over-tolerant” latent factors directly correspond to the decrease of the model’s description sources. These problems reduce the DBN’s description ability as well as the classification performance. The first problem is solved by trying to perform the latent factors diversely. The “dead” and “potential over-tolerant” latent factors (neurons) are related to the sparsity and selectivity of activations of visual neurons and the selectivity and sparsity are just two epiphenomena of the diversity of receptive fields. Hence, both the problems can be solved together by diversifying the DBN models.

The classification performance enhancement through the diversification of latent factors of a given model has become attractive topic in recent years [66–68]. The determinantal point process (DPP) is used as a prior for probabilistic latent variable models in [68]. Probabilistic latent variable models are one of the vital elements of machine learning. The determinantal point process enables a modeler to specify a notion of similarity on the space of interest, which in this case is a space of possible latent distributions, via a positive definite kernel. The DPP then assigns probabilities to particular configurations of these distributions according to the determinant of the Gram matrix. This construction naturally leads to a generative latent variable model in which diverse sets of latent parameters are preferred over redundant sets.

Figure 4. The simple structure of the standard DBN. (RBM- Restricted Boltzmann Machine).
Restricted Boltzmann Machine (RBM)s has demonstrate immense effectiveness in clustering and classification. In [69], diversified RBM (DRDM) is proposed to enhance the diversity of the hidden units in RBM. To combat the phenomenon that many redundant hidden units are learned to characterize the dominant topics as best as possible with the price of ignoring long-tail topics by imposing a diversity regularizer over these hidden units to reduce their redundancy and improve their coverage of long-tail topics. First-order Hidden Markov Models (HMM) provides a fundamental approach for unsupervised sequential labeling. A diversity-encouraging prior over transition distributions is incorporated to extend HMM to diversified HMM (dHMM) [66]. The dHMM shows great effectiveness in both the unsupervised and supervised settings of sequential labeling problems. A successful attempt has been made to improve the HSI classification by diversifying a deep model in [70]. A new diversified DBN is developed through regularizing pretraining and fine-tuning procedures by a diversity promoting prior over latent factors. Moreover, the regularized pretraining and fine-tuning can be efficiently implemented through usual recursive greedy and back-propagation learning framework.

The conventional applications of the diversified models include image classification [69], image restoration [67], and video summarization [71].

Two hyperspectral data sets, Indian Pines and the University of Pavia scenes are selected for the evaluation of diversified DBN (D-DBN)-based classification method. The Indian Pines data set has 220 spectral channels in 0.4 to 2.45 μm region of the visible and infrared spectrum with a spatial resolution of 20 m × 20 m. The 20 spectral bands were removed due to noise and water absorption, and the data set contains 200 bands of size 145 × 145 pixels. A three-band false color image and the ground truth data are presented in Figure 5. The University of Pavia data set with a spectral coverage ranging from 0.43 to 0.86 μm is presented in Figure 6. The image contains 610 × 340 pixels and 115 bands. After removing 12 bands due to noise and water absorption, the image contains 103 bands with a spatial resolution as 1.3 m × 1.3 m.

The structure of the DBN for the Indian Pines data set is set as 200 – 50 - ... - 50 - 8, which means the input layer has 200 nodes corresponding to the dimension of input data, the output layer has eight nodes corresponding to the number of classes, and all the middle layers have 50 nodes. Particulars about the number of training and testing samples are presented in Table 1. The performance of the DBN can be significantly improved by modifying the pretraining and fine-tuning of D-DBNs. DBN-based classification methods realizes comparatively fast inference and

![Figure 5](image)

*Figure 5.* Indian Pines data set. (a) Original image produced by the mixture of three bands. (b) Ground truth with eight classes. (c) Map color.
7.2 Convolutional neural networks (CNN)

Quite a few number of neural network-based classification methods have been proposed in the literature to deal with both supervised and unsupervised nonparametric approaches [72–74]. The feedforward neural network (FN)-based classifiers are extensively used with the variation of second-order optimization-based strategies, which are faster and need fewer input parameters [75, 76]. The extreme learning machine (ELM) learning algorithm has become popular that train single hidden-layer FNs (SLFN) [77, 78]. Then, the concept has been extended to multi-hidden-layer networks [79], radial basis function (RBF) networks [80], and kernel

competent representation of hyperspectral image and thus good classification performance.
learning [81, 82]. ELM-based networks are remarkably efficient in terms of accuracy and computational complexity and have been successfully applied as nonlinear classifiers for hyperspectral data, providing results comparable with state-of-the-art methodologies.

In recent years, convolutional neural network (CNN) has acquired auspicious achievements in remote sensing [58, 83–85]. The deep structure of CNNs allows the model to learn highly abstract feature detectors and to map the input features into representations that can clearly boost the performance of the subsequent classifiers. The advantage of such approaches over probabilistic methods result mainly from the fact that neural networks do not need prior knowledge about the statistical distribution of the classes. Their attractiveness increased because of the availability of feasible training techniques for nonlinearly separable data citepbenediktsson1990statistical, although their use has been traditionally affected by their algorithmic and training complexity [86] as well as by the number of parameters that need to be tuned.

The CNN is a multi-layer architecture with multiple stages for effective feature-extraction. Generally, each stage of CNN is composed of three layers. (i) convolutional layer (ii) nonlinearity layer and (iii) pooling layer. The classical CNN is composed of one, two, or three feature-extraction stages, followed by one or more fully connected layers and a final classifier layer.

Convolutional layer: The input to the convolutional layer is represented as \( x_{i,m,n} \), with \( r \) number of features maps \( x^i \), each map is of size \( m \times n \). The convolutional layer consists of filter banks \( W \) of size \( l \times l \times q \) that connects input filter map to output filter map. The output of convolutional layer is a three-dimensional array \( m_1 \times n_1 \times k \), composed of \( k \) feature maps of size \( m_1 \times n_1 \). The output of the convolutional layer is determined as:

\[
 z^s = \sum_{i=1}^{q} W^s_i \ast x^i + b_s 
\]  

(32)

Where, \( b \) is the bias parameter.

Nonlinearity layer: The nonlinearity layer measures the output feature map \( a^s = f(z^s) \), as \( f(.) \) is usually selected to be a rectified linear unit (ReLU) \( f(x) = \max(0,x) \).

Pooling Layer: The pooling layer involves executing a max operation over the activations within a small spatial region G of each feature map: \( p^s_G = \max_{i \in G} a_i^s \).

After the multiple feature-extraction stages, the entire network is trained with back propagation of a supervised loss function such as the classic least-squares output, and the target output \( \gamma \) is represented as a L-of-K vector, where K is the number of output and L is the number of layers:

\[
 J(\theta) = \sum_{i=1}^{N} \left( \frac{1}{2} \| h(x_i, \theta) - \gamma \|^2 \right) + \lambda \sum_{l=1}^{L} \text{sum} \left( \| \theta^{(l)} \|^2 \right), 
\]  

(33)

where \( l \) indexes the layer number. Primary goal is to minimize \( J(\theta) \) as a function of \( \theta \). To train the CNN, stochastic gradient descent with back propagation is exercised to optimize the function.

The three fundamental parts of a CNN are a convolutional layer, non linear function and a pooling layer. A deep CNN can be formulated by stacking several convolution layers with nonlinear operation and several pooling layers. A deep CNN can hierarchically extract the features of inputs, which tend to be invariant and robust [87]. The architecture of a deep CNN for spectral classification is shown in Figure 7.

Processing and Analysis of Hyperspectral Data
A systematic survey on deep networks for remote sensing data has been presented in [56]. In [83], CNN was investigated to exploit deep representation based on spectral signatures and the performance proved to be superior to that of SVM. The high level spatial features are extracted using CNN [88], deep CNN for pixel classification while learning unsupervised sparse features [59], deep CNN to learn pixel-pair features [89] and few more.

The performance of the HSI classification method proposed in [83] termed as deep CNN (D-CNN) is compared with a traditional SVM classifier. Two hyperspectral data sets including Indian Pines and University Of Pavia are used for the evaluation. The Indian Pines data set consists of 220 spectral channels in the 0.4–2.45 μm region of the visible and infrared spectrum with a spatial resolution of 20 m. The University of Pavia data set with a spatial coverage of 610 × 340 pixels covering the city of Pavia and has 103 spectral bands prior to water band removal. It has a spectral coverage from 0.43 to 0.86 μm and a spatial resolution of 1.3 m. All the layer parameters of these two data sets for CNN classifier are set as specified in [83]. The comparison of classification performance between D-CNN and SVM is presented in Table 2. Figures 8 and 9 interpret the corresponding classification.

<table>
<thead>
<tr>
<th>Data set</th>
<th>D-CNN (%)</th>
<th>SVM (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indian pines</td>
<td>90.18</td>
<td>87.54</td>
</tr>
<tr>
<td>University of Pavia</td>
<td>92.64</td>
<td>90.42</td>
</tr>
</tbody>
</table>

Table 2. Comparison of results between the D-CNN and SVM using two data sets.
maps obtained with D-CNN and SVM classifier. Furthermore, compared with traditional SVM the D-CNN classifier has higher classification accuracy for the overall data sets.

Furthermore, the application of Deep learning to hyperspectral image classification has some potential issues to be investigated.

i. Deep learning methods may lead to a serious problem called overfitting, which means that the results can be very good on the training data but poor on the test data. To deal with this issue, it is necessary to use powerful regularization methods.

ii. In contrast to natural images, the high resolution remote sensing (RS) images are complex in nature. The complexity of RS images leads to some difficulty in discriminative representation and learning features from the objects with DL.

iii. The deeper layers in supervised networks like CNNs can learn more complex distributions. Research on appropriate depth for a DL model for a given data set is still an open research topic to be explored.

iv. Deep learning methods can be combined with other methods, such as sparse coding and ensemble learning which is another research area in hyperspectral data classification.

Figure 9. Thematic maps resulting from classification for University of Pavia data set. From left to right: ground truth, SVM, and D-CNN.
References


[66] Qiao M, Bian W, Da Xu RY, Tao D. Diversified hidden Markov models for sequential labeling. IEEE Transactions on Knowledge and Data Engineering. 2015;27(11):2947-2960


[81] Huang G-B. An insight into extreme learning machines: Random neurons,


