Journal of Geovisualization and Spatial Analysis (2020) 4:13
https://doi.org/10.1007/s41651-020-00048-5

REVIEW

Advances of Four Machine Learning Methods for Spatial Data Handling: a Review

Peijun Du[1,2,3] · Xuyu Bai[1,2,3] · Kun Tan[4] · Zhaohui Xue[5] · Alim Samat[6] · Junshi Xia[7] · Erzhu Li[8] · Hongjun Su[5] · Wei Liu[8]

Abstract

Most machine learning tasks can be categorized into classification or regression problems. Regression and classification models are normally used to extract useful geographic information from observed or measured spatial data, such as land cover classification, spatial interpolation, and quantitative parameter retrieval. This paper reviews the progress of four advanced machine learning methods for spatial data handling, namely, support vector machine (SVM)-based kernel learning, semi-supervised and active learning, ensemble learning, and deep learning. These four machine learning modes are representative because they improve learning performances from different views, for example, feature space transform and decision function (SVM), optimized uses of samples (semi-supervised and active learning), and enhanced learning models and capabilities (ensemble learning and deep learning). For spatial data handling via machine learning that can be improved by the four machine learning models, three key elements are learning algorithms, training samples, and input features. To apply machine learning methods to spatial data handling successfully, a four-level strategy is suggested: experimenting and evaluating the applicability, extending the algorithms by embedding spatial properties, optimizing the parameters for better performance, and enhancing the algorithm by multiple means. Firstly, the advances of SVM are reviewed to demonstrate the merits of novel machine learning methods for spatial data, running the line from direct use and comparison with traditional classifiers, and then targeted improvements to address multiple class problems, to optimize parameters of SVM, and to use spatial and spectral features. To overcome the limits of small-size training samples, semi-supervised learning and active learning methods are then utilized to deal with insufficient labeled samples, showing the potential of learning from small-size training samples. Furthermore, considering the poor generalization capacity and instability of machine learning algorithms, ensemble learning is introduced to integrate the advantages of multiple learners and to enhance the generalization capacity. The typical research lines, including the combination of multiple classifiers, advanced ensemble classifiers, and spatial interpolation, are presented. Finally, deep learning, one of the most popular branches of machine learning, is reviewed with specific examples for scene classification and urban structural type recognition from high-resolution remote sensing images. By this review, it can be concluded that machine learning methods are very effective for spatial data handling and have wide application potential in the big data era.

Keywords Machine learning · Remote sensing image classification · Spatial interpolation · Support vector machine · Ensemble learning · Deep learning · Semi-supervised learning · Active learning

✉ Peijun Du
peijun@nju.edu.cn

1 School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China

2 Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources of China, Nanjing 210023, China

3 Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

4 Key Laboratory of Geographic Information Science, Ministry of Education, East China Normal University, Shanghai 200241, China

5 School of Earth Sciences and Engineering, Hohai University, Nanjing 211100, China

6 State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China

7 RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan

8 School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou 221116, China

Published online: 14 May 2020

Introduction

With the rapid progress of spatial data acquisition technologies such as in remote sensing and geo-sensor networks, more and more spatial data with image or point formats at different resolution have been collected. Big Earth Data, or similar terms including spatial big data, geographic big data, or spatio-temporal big data, has become a new direction of multidisciplinary integration (Guo et al. 2017). How to derive useful geographic and thematic information on a wide range from limited observed or measured sample data is an essential challenge for geoinformatics and geospatial-related applications. Spatial data handling, for example, spatial interpolation, land cover classification, and quantitative parameter retrieval, play important roles in extracting useful geographic information from observed or measured data (Li et al. 2008; Shi et al. 2012). Traditional mathematical and statistical methods or models have been widely used for spatial data handling, and geographic laws and constraints have been introduced into statistical models, resulting in some new methods for spatial data processing, for example, geographically weighted regression (GWR) (Fotheringham et al. 2015) and geographically and temporally weighted regression (GTWR) (Huang et al. 2010a; Fotheringham et al. 2015).

Machine learning (ML) is an essential theoretic/technical support for spatial data handling in the era of artificial intelligence and big data. Most of the machine learning tasks can be categorized into two groups: classification and regression (Bishop 2006; Ethem 2010). This is true for spatial data handling as well. According to Kanevski et al. (2008, 2009), the important questions for spatial data analysis and modeling consist of the development of data-adaptive, nonlinear, robust, and multivariate models in high-dimensional spaces with functional generalization capacity. These requirements actually fall in the scope of machine learning.

Based on the common points of spatial data handling and machine learning, that is, deriving or recognizing information on the whole domain based on limited training samples and specific algorithms, this paper reviews the progress of some advanced machine learning methods for spatial data handling. We mainly focus on the spatial data analysis and handling tasks that can be described by classification and regression models. For spatial data handling via machine learning, three key elements are learning algorithm, training samples, and input features that can be improved by the four machine learning models. Four representative machine learning methods are reviewed for spatial data handling in this paper: support vector machine (SVM) as an effective kernel learning algorithm based on structural risk minimization criterion, semi-supervised and active learning to solve small-size training sample problems, ensemble learning to integrate different algorithms, and deep learning to learn deep features hidden in the data. The advantages of these four machine learning methods correspond to the three elements in spatial data handling.

In order to apply machine learning methods to spatial data handling successfully, a four-level strategy is adopted, respectively: (1) experimenting and evaluating the applicability, (2) extending the algorithms by embedding spatial properties, (3) optimizing the parameters for better performance, and (4) enhancing the algorithm by multiple means. Table 1 presents the structure of this review according to the spatial data handling tasks and machine learning methods based on this four-level strategy. Most of the examples come from the authors' work in this field but we do summarize some representative publications from other scholars.

After a brief introduction, "Support Vector Machine for Classification and Regression: Improvements and Optimization" summarizes the advances of support vector machine (SVM) to demonstrate the merits of novel machine learning algorithms for spatial data. Semi-supervised learning and active learning with insufficient labeled samples are reviewed in "Semi-supervised and Active Learning for Classification with Small-Size Training Samples." Furthermore, ensemble learning is employed to integrate the advantages of multiple learners and enhance the generalization capacity in "Ensemble Learning for Improving Unstable Algorithms." Deep learning is used for scene classification and urban structural type recognition from high-resolution remote sensing images in "Deep Learning for Scene Classification and Urban Structural Type Recognition." Finally, "Conclusions and Prospects" draws the conclusions and hints some possible perspectives.

Support Vector Machine for Classification and Regression: Improvements and Optimization

Direct Use of SVM for Remote Sensing Image Classification

SVM is a binary classifier based on Vapnik-Chervonenkis (VC) dimension theory and minimum structural risk criterion (Cortes and Vapnik 1995) that is utilized to implicitly map the raw data into a very high-dimensional space for better separation using a quadratic optimization approach and kernel function. The basic problem of binary SVM classification is finding a decision boundary in the kernel space that can maximize the margins of the decision hyperplane (Boser et al. 1992; Cortes and Vapnik 1995). The boundary conditions can be searched using a typical convex optimization problem, which is usually solved by the sequential minimal optimization (SMO) algorithm (Platt et al. 2000). SMO algorithm selects and updates a pair of Lagrange multipliers in each iteration which meets the Karush-Kuhn-Tucker (KKT) condition.

Table 1 The structure and main content of this review paper

Four-level strategy	Classification				Regression	
	SVM	SSL/AL	EL	DL	SVM	EL
Evaluating applicability	SVM and multi-class SVM	SSL/AL for small size	Rotation forest	Scene classification	Biophysical parameter retrieval	Ensemble interpolation
Embedding spatial properties	Spatial-spectral SVM	Spatial-spectral SSL	Spatial-spectral RoF	Multi-scale DL	SVM with Kriging	Spatial partitioning ensemble
Optimizing parameters	PSO SVM	–	DCS-MCS	–	PSO SVM LSSVM	–
Enhancing performance	Multi-kernel SVM rotation-based SVM	Caps-TripleGAN SSL	Ensemble ELM	Multi-model DL	Multi-kernel regression	–

SVM support vector machine, *SSL* semi-supervised learning, *AL* active learning, *EL* ensemble learning, *DL* deep learning, *MCS* multiple classifier system, *PSO* particle swarm optimization, *DCS* dynamic classifier selection, *LSSVM* least square support vector machine

To accelerate the convergence of SMO, Keerthi et al. (2001) and Fan et al. (2005) improved the first-order and second-order heuristic learning process. Compared with other statistical learning algorithms, SVM has the advantage on be less dependent on prior knowledge, greater suitability for small sample sizes (Chi et al. 2008), greater robustness to noise (Du et al. 2012a), more powerful generalization capacity, and higher learning efficiency (Ancona et al. 2006).

SVM has been widely used for spatial data processing, mostly in hyperspectral remote sensing image classification (Melgani and Bruzzone 2004), object detection (Inglada 2007), and feature selection (Pal and Foody 2010). Mountrakis et al. (2011) reviewed the contributions of SVM in the field of remote sensing and concluded that SVM is a reliable method in the remote sensing community. SVM has been verified to separate heterogeneous classes with a very limited number of training set (Hermes et al. 1999). Roli and Fumera (2001) presented the application of the SVM on the multi-sensor image classification and discussed the design and the training phase of the SVM. Experimental results revealed the advantages of SVM over multi-layer perceptron (MLP) neural networks and k-NN classifiers. Huang et al. (2002) compared the performances of the SVM and three other classifiers in land cover classification, and they showed that kernel type and parameters affect the decision boundary shape and thus influence the performance of SVM. The high accuracies can be ascribed to the capacity of locating an optimal separation hyperplane.

The properties of the SVM in hyper-dimensional spaces were evaluated, and the results showed that SVM exhibits lower sensitivity to the Hughes phenomenon, which brings an excellent solution to avoid the time-consuming phase in feature reduction (Melgani and Bruzzone 2004). Moreover, SVM has made a breakthrough in the field of object detection using high-resolution remote sensing images (Inglada 2007). The SVM-based detection method is promoted by geometric features that describe different geometric properties of several classes. Pal and Foody (2010) found SVM is sensitive to the number of features which would encounter the Hughes phenomenon when training on a small-size dataset. Hence, a feature selection operation is recommended to include with respect to kernel function before a classification analysis, in which the between-class and within-class information was utilized to tackle the Hughes phenomenon (Kuo et al. 2013).

Improving SVM for Multi-class Classification Tasks

The two-class classification of traditional SVM brings difficulties while handling spatial data with multiple classes. Therefore, it is of great significance to extend the traditional algorithm to multi-class cases. Multi-class strategies can be grouped into two types: one is constructing and combing several binary SVM classifiers and the other is considering all data in one optimization formulation directly (Hsu and Lin 2002). Here we introduce four popular strategies.

1-Against-All (1-a-a) Strategy N binary SVM classifiers are constructed in this strategy, each classifier corresponding to one class. Testing samples are split into N two-class classifiers. The discriminant function of each classifier is carried out, then the class with the largest discriminant function for testing data is chosen as the result.

1-Against-1 (1-a-1) Strategy Each of the two classes needs to be processed by a child SVM classifier, that is, $N \times (N-1)/2$ SVM classifiers are required. By combining all classifiers, SVM can determine the class of each pixel through the accumulation of predicting a classification. The disadvantage of this strategy is that the number of support vectors grows super linearly with the number of classes.

Decision Directed Acyclic Graph (DDAG) Strategy A DAG is a finite directed graph whose edges have an orientation and no directed cycles. The DDAG can be implemented using the decision nodes obtained by $N \times (N-1)/2$ child classifiers of SVM, where each node excepts one class from all classes. If the node chose the one class, the other is eliminated. A test class is evaluated against the decision node that corresponds to the first and last classes of the list. Therefore, in order to separate two subgroups, more support vectors are needed, which increases the time consumption of the training and testing process.

Binary Tree Strategy Binary tree classifiers decompose a complex problem into the sub-problems. The original multi-class is divided into multiple two-class nodes. The binary tree strategy-based SVM has two representative structures: slant binary tree SVM (SBTSVM) and the balanced binary tree SVM (BBTSVM).

A novel binary tree SVM classifier based on the properties of Jeffries-Matusita (J-M) distance was defined (Du et al. 2012a). In the strategy, the most easily separated classes are classified firstly, and the derived separability measures are introduced into the decision tree at the construction phase. This binary tree SVM (BTSVM) includes two improved algorithms. One is the adaptive binary tree SVM (ABTSVM) based on J-M distance, and the other is Kullback-Leibler distance binary tree SVM (KLBTSVM), which uses Kullback-Leibler distance instead.

Wavelet Kernel and Composite Kernels

SVM is confronted with the bottleneck of kernel parameter selection. Different kernels have significant differences in time consumption and classification accuracy. There are a variety of contributions to the wavelet kernel function for SVM, such as wavelet SVM (WSVM) (Zhang et al. 2004), reproducing the wavelet kernel frame (Rakotomamonjy et al. 2005) and least square SVM (Wu and Zhao 2006).

However, most of these studies focused on support vector regression, and the applications in hyperspectral classification are not yet mature. To solve this problem, Du and Tan (2010) proposed a method combining SVM with wavelet analysis to construct a WSVM classifier, which is based on wavelet kernel functions in reproducing kernel Hilbert space (RKHS). The wavelet kernel in RKHS is one kind of multidimensional wavelet function that can approximate arbitrary nonlinear functions. The significance of semiparametric estimation is proposed in this work. Experiments with ROSIS data were carried out using Haar, Daubechies, Coiflets, and Symlets as wavelet kernel SVM function, respectively. And the results reveal that WSVM is more effective than the RBF kernel SVM, SAM, and MDC. The Coiflet kernel WSM obtains the best performance, which possesses the maximal number of vanishing shifted scaling moments.

To address the combination of spectral-spatial features in hyperspectral images, Marconcini et al. (2009) defined a novel composite classifier based on SVM, whose composite kernel functions effectively considered spectral information and spatial content. Tan and Du (2011) proposed a method of applying SVM classifiers with different kernels. Experimental results show that adopting more sophisticated cross-kernel approaches can obtain better performance, and the wavelet texture has the potential on hyperspectral image classification.

Fang et al. (2015) adopted the superpixel and a fixed-size region to present SC-MK algorithm, in which the size and shape of the superpixel can be adjusted adaptively according to the spatial structures. In addition, multiple kernels are utilized in SC-MK for spectral-spatial feature exploitation within and among superpixels. Experimental results reveal the superiority in terms of both quantitative metrics and visual quality on the classification map.

Gu et al. (2016) proposed a multiple kernel learning (MKL) framework to incorporate spectral and spatial features for hyperspectral image classification, namely multiple-structure-element nonlinear MKL (MultiSE-NMKL). The combination of MultiSE and base kernels can explore the similarity information generated by different kernels, which is of great significance to improve the discriminability.

Wang et al. (2016) proposed a discriminative multiple kernel learning (DMKL) method for spectral image classification, which aims at learning the optimal combined kernel by maximizing separability in reproduction kernel Hilbert space. DMKL achieves the maximum separability by finding the best projective direction. One of DMKL's variants conducts Fisher criterion (FC) to find the optimal projective direction, namely DMKLFC, and the other uses maximum margin criterion (MMC), titled DMKLMMC. All the basic kernels are projected to form a distinct combined kernel after learning the projective direction, which has the following advantages: (1) improve the classification performance without any constraints for basic kernel selection; (2) the competitive scale of spatial filters can be selected by sorting the corresponding weights; and (3) the computational burden is reduced by fewer support vectors.

Xia et al. 2015proposed an ensemble method, rotation-based SVM (RoSVM) that combines SVMs and multiple classifier systems (MCSs). Diverse SVM classification results are generated using random feature selection and data transformation in RoSVM, in which the hyper-parameters are chosen carefully for different datasets.

Tan and Du (2010) explored the accuracy of multi-kernel SVM with morphological profiles, which is carried out to obtain several principal components (PCs) from the hyperspectral data, then the morphological profile is established for each of the PCs and is integrated as one morphological profile.

Moreover, some hybrid machine learning methods have been explored, such as the spatial-spectral feature-based hybrid model of convolutional neural network (CNN) and SVM (Leng et al. 2016). In summary, multi-kernel methods succeeded in taking advantage of both the spatial and spectral information.

Table 2 summarizes the overall accuracy (OA) and kappa coefficients of some typical multiple kernel SVMs for remote sensing images.

Optimizing SVM Parameters by Particle Swarm Optimization (PSO)

SVM has a tendency to over-fit since one needs to determine the involved kernel and penalty parameter, greatly limiting the applications of SVM. Therefore, cross-validation (CV) (Hastie et al. 2009), genetic algorithm (GA) (Frohlich et al. 2003), and PSO (Kennedy et al. 2001) have been widely used for parameter optimizing of SVM. PSO is a biological intelligent algorithm, in which the particle adjusts its speed and direction following the optimal particle. Relevant research have proved that PSO is a good choice to enhance the performance of SVM for hyperspectral image classification (Xue et al. 2014, 2015).

Table 3 reports the classification accuracies obtained by different features and classifiers on the University of Pavia ROSIS hyperspectral image. According to the table, the proposed harmonic analysis based on SVM optimized by PSO (HA-PSO-SVM) method obtains the highest overall accuracy, with an OA of 93.4%. Figure 1 shows the classification maps, the "bare soil" region presents a great variation among these classification maps obtained by different methods, which caused by the fact that the region is highly mixed by various surface objects.

Spatial-Spectral Classification by SVM

Spectral-spatial classification has been well developed in the remote sensing community during the last decade (Fauvel et al. 2012). Advances of the spectral-spatial classification are triggered by such methods as structural filtering (Camps-Valls et al. 2006), random field (Zhong and Wang 2010; Li et al. 2011b), image segmentation (Tarabalka et al. 2010a, b), mathematical morphology (Fauvel et al. 2008; Dalla Mura et al. 2010; Ghamisi et al. 2016), sparse representation (Chen et al. 2011), and deep convolutional neural networks (Romero et al. 2015). Those approaches revealed the powerful potential of combining spatial and spectral information for remote sensing image classification.

Among many spectral-spatial classification methods, integrating image separation with classification is a new trend in the literature, where images can be represented with piecewise smooth (content) and texture components by the exhaustively sparse representation-based splitting scheme. A morphological component analysis (MCA)-based image separation approach was proposed by Starck et al. (2005), and the separated components are represented by the transformed coefficients. Xue et al. (2015) designed a new scheme for spectral-spatial classification of hyperspectral images. In the proposed method, a sparse representation is introduced into MCA. Moreover, the Curvelet and Gabor transform are carried out to generate the content and texture dictionaries. Then, a sparse unmixing by variable splitting and augmented Lagrangian (SUnSAL) algorithm is adopted for the image separation based on sparse representation. To reduce the computational complexity and retain the spectral information, dimensionality reduction is performed before applying MCA on the input data. The proposed approach for spatial and spectral feature extraction is then combined with an SVM classifier.

Table 4 reports the classification accuracies obtained by different spatial-spectral classification methods. As reported in Table 4, the proposed morphological component analysis-based image separation (MCASUnSAL) integrated with MNF features and the SVM classifier (M^2S^3VM) obtained the highest classification accuracy with an OA of 99.01%, which is 5% higher than the baseline method. Classification maps are shown in Fig. 2, and M^2S^3VM produced more smooth and accurate classification results compared with other counterparts.

Support Vector Machine for Regression and Spatial Interpolation

In addition to remote sensing image classification, kernel-based SVM models are also widely used in quantitative parameter retrieval from soil spectroscopy (Vohland et al. 2011),

Table 2 The OAs and kappa coefficients obtained by SVM using different kernels

Methods	OA	Kappa
multi-SVM (Tan and Du 2010) (10% of the labeled samples)	91.05	0.9027
SC-MK (Fang et al. 2015) (200 samples/class)	99.22	0.99
RoSVM-RPB (Xia et al. 2015) (10% of the labeled samples)	77.53	–
MultiSE-NMKL (Gu et al. 2016) (3% training samples)	95.16	–
DMKL-FC (Wang et al. 2016) (50 samples/class)	97.13	–
Xue's method (Xue et al. 2015) (5% of the labeled samples)	99.01	0.987

Table 3 A comparison of different optimization methods for the ROSIS University of Pavia dataset (Xue et al. 2014)

Class	PSO-SVM					CV-SVM					GA-SVM				
	HA	MNF	PCA	ICA	RAW	HA	MNF	PCA	ICA	RAW	HA	MNF	PCA	ICA	RAW
OA	*93.4 ± 0.2*	92.1 ± 0.2	91.0 ± 0.1	89.1 ± 0.4	93.1 ± 0.3	93.3 ± 0.2	92.2 ± 0.2	91.0 ± 0.1	89.1 ± 0.4	93.2 ± 0.2	91.8 ± 2.0	91.2 ± 1.7	90.3 ± 1.1	86.6 ± 1.8	90.9 ± 1.7
AA	91.4 ± 0.3	90 ± 0.3	88.7 ± 0.3	86.8 ± 0.6	91.5 ± 0.3	91.2 ± 0.3	90.1 ± 0.4	88.8 ± 0.2	86.8 ± 0.6	*91.6 ± 0.2*	89.6 ± 2.2	88.8 ± 2.3	88.0 ± 1.4	81.9 ± 2.9	89.1 ± 1.8
κ	*0.91 ± 0*	0.89 ± 0	0.88 ± 0	0.85 ± 0	*0.91 ± 0*	*0.91 ± 0*	0.90 ± 0	0.88 ± 0	0.85 ± 0	*0.91 ± 0*	0.89 ± 0	0.88 ± 0	0.87 ± 0	0.82 ± 0	0.88 ± 0

which can effectively solve the nonlinear problem with high performance. At the same time, in the field of predicting soil organic matter content and clay content, SVM also shows great potential (Rossel and Behrens 2010; Nawar et al. 2016). Tan et al. (2018) presented an improved estimation model, CARS-PLS-SVM, to tackle the nonlinear problem in multiple sites based on SVM. The proposed model can be applied to both feature extraction and small sample learning, with obvious reduction of computational load and field work costs.

A review of support vector machine regression (also namely support vector regression, SVR) for quantitative retrieval and parameter estimation from remote sensing data was conducted by Mountrakis et al. (2011), covering the retrieval of biophysical parameter, chlorophyll concentration, evapotranspiration, land and sea surface temperatures, ocean primary productivity by using general support vector machine, relevance vector machine as a variant of SVMs, automatic parameter optimization method for SVM regression, and least squares SVM (LS-SVM).

Li et al. (2011a) analyzed the applications of machine learning methods to spatial interpolation of environmental variables, mainly compared the performance of SVM and random forest (RF) with traditional spatial interpolation methods including ordinary kriging (OK) and inverse distance squared (IDS), and combined them together to form SVMOK and SVMIDS for mud content estimation. The study has opened an alternative source of methods for spatial interpolation of environmental properties.

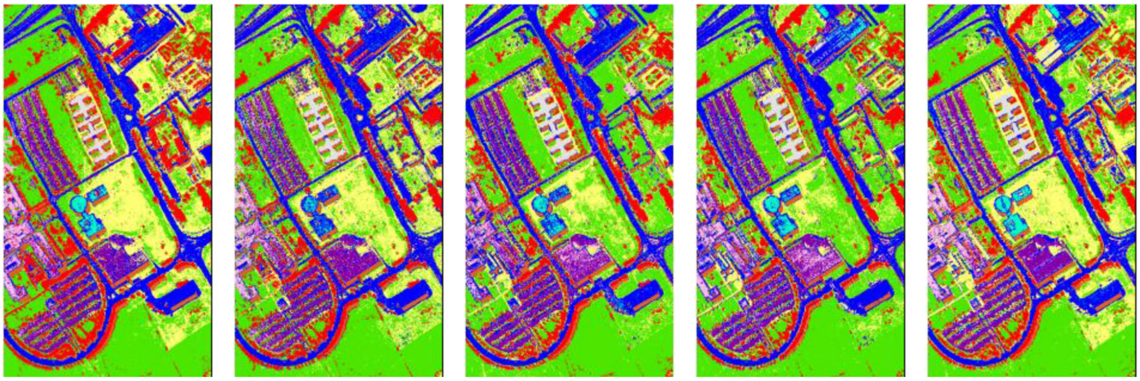
In terms of spatial interpolation, the models of support vector regression can also be used for DSM generation from point could. For example, Shi et al. (2009) used LSSVM approach to generate digital surface model (DSM) from LiDAR point could data, and LS-SVM was concluded to be more effective in terms of noise reduction, computational efficiency, and accuracy in DSM generation.

Semi-supervised and Active Learning for Classification with Small-Size Training Samples

Spatial data handling is often confronted with the challenges of limited, even insufficient labeled training samples, and semi-supervised and active learning strategies are quite effective in tackling small-size training sample problems.

Semi-supervised Learning for Hyperspectral Image Classification

Semi-supervised learning originated from self-training, a classification algorithm based on unlabeled samples proposed by Scudder (1965). With the development of natural language



(a) HA (93.4%) (b) MNF (92.1%) (c) PCA (91.0%) (d) ICA (89.1%) (e) RAW (93.1%)

Fig. 1 Classification maps obtained by PSO optimized SVM for the ROSIS University of Pavia dataset based on different input features (Xue et al. 2014). **a** HA (93.4%). **b** MNF (92.1%). **c** PCA (91.0%). **d** ICA (89.1%). **e** RAW (93.1%)

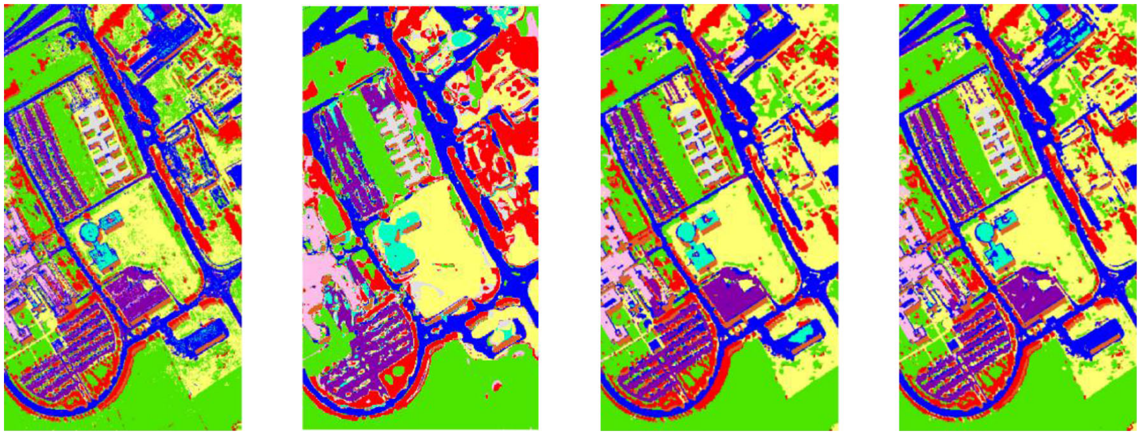
processing, semi-supervised learning has become a research hotspot in machine learning (Zhu and Goldberg 2009). Standard semi-supervised learning algorithms include multi-view learning (Yu et al. 2012; Gulp et al. 2009), self-learning (Dópido et al. 2013; Tuia et al. 2009), co-training (Samiappan and Moorhead 2015; Blum and Mitchell 1998), graph-based approaches (Bai et al. 2012), and transductive support vector machines (TSVM) (Chen et al. 2003).

Semi-supervised learning has recently attracted considerable attention in the field of hyperspectral remote sensing image analysis. Dópido et al. (2012) exploited active learning for unlabeled samples' selection. Wang et al. (2015) proposed a methodology on the basis of spatial-spectral label propagation. Xia et al. (2014b) extended semi-supervised probabilistic principal component analysis, semi-supervised local fisher discriminant analysis, and semi-supervised dimensionality reduction with pairwise constraints to hyperspectral image for feature extraction. In Tan's (2015) research, class labels of selected unlabeled samples are determined by combining spatial neighborhood information.

Meanwhile, Blum and Mitchell (1998) developed co-training, which has been widely used in semi-supervised learning. This algorithm requires two sufficient and redundant views, which cannot be achieved for hyperspectral imagery. Then, Goldman and Zhou (2000) proposed statistical co-training, employing two different learning algorithms on the basis of a single view. Zhou and Goldman (2004) proposed democratic co-training that contributes to label the selected unlabeled samples and to produce the final hypothesis using a cross-validation technique which is quite time-consuming. To solve

Table 4 Accuracies obtained by different spectral-spatial classification methods for the ROSIS University of Pavia dataset (Xue et al. 2015)

Class	No. of samples Train	Classification methods Test	SVM	M^2S^3VM	LORSAL-MLL	MPM-LBP
Asphalt	332	6299	94.51 ± 0.14	98.51 ± 0.33	*98.03 ± 0.52*	98.13 ± 0.73
Meadows	932	17,717	97.17 ± 0.16	99.81 ± 0.08	*99.89 ± 0.06*	99.74 ± 0.17
Gravel	105	1994	82.94 ± 0.88	*96.52 ± 0.83*	79.18 ± 6.29	81.57 ± 1.94
Trees	153	2911	96.12 ±	*97.88 ± 0.31*	96.11 ± 0.97	96.56 ± 1.11
Painted metal sheets	67	1278	99.41 ± 0.23	*99.92 ± 0.03*	99.30 ± 0.33	99.06 ± 0.42
Bare soil	251	4778	91.06 ± 0.48	*99.88 ± 0.09*	99.53 ± 0.27	98.23 ± 1.01
Bitumen	67	1263	87.26 ± 0.63	*99.09 ± 0.13*	92.07 ± 4.12	92.92 ± 2.64
Self-blocking bricks	184	3498	88.32 ± 0.61	96.74 ± 0.84	95.45 ± 0.94	*96.77 ± 0.61*
Shadows	47	900	*99.94 ± 0.04*	97.70 ± 0.74	99.71 ± 0.27	99.84 ± 0.10
Average accuracy	–	–	92.97 ± 0.19	*98.53 ± 0.13*	95.56 ± 1.15	95.56 ± 0.41
Overall accuracy	–	–	94.33 ± 0.19	*99.01 ± 0.14*	97.75 ± 0.48	97.71 ± 0.20
k statistic	–	–	0.925 ± 0.003	*0.987 ± 0.002*	0.970 ± 0.006	0.970 ± 0.003
Time (s)	–	–	437.96	2568.81	*35.75*	1084.94



(a) SVM (94.33%) (b) **M²S³VM** (99.01%) (c) LORSAL-MLL (97.75%) (d) MPM-LBP (97.71%)

Fig. 2 Classification maps obtained by different spectral-spatial classification methods for the ROSIS University of Pavia dataset (Xue et al. 2015)

this problem, Zhou and Li (2005) developed tri-training which neither requires the instance space to be described with sufficient and redundant views nor imposes any constraints on supervised learning algorithms, and has broader applicability compared with previous co-training algorithms.

Tan et al. (2014) developed a semi-supervised SVM for the hyperspectral image classification, namely S2SVMSE, which use the segmentation algorithm to extract spatial information for unlabeled sample selection. In this work, the unlabeled samples are the most similar to the labeled ones, and then the candidate set of unlabeled samples is enlarged by utilizing the mean shift-based segmentation result. Tan et al. (2016) further proposed a semi-supervised tri-training algorithm, namely TT_AL_MSH_MKE, in which three measures of diversity, i.e., the disagreement metric, the double-fault measure, and the correlation coefficient, are carried out for the combination of the optimal classifiers. An active learning method is utilized to select unlabeled samples. Experimental results with the ROSIS data demonstrate that TT_AL_MSH_MKE has satisfactory performance.

Recently, based on regularized local discriminant embedding (RLDE), Ou et al. (2019) proposed a novel tri-training algorithm. To solve the problems of over-fitting and singular values, RLDE is used for the extraction of optimal features. Then, an active learning process is carried out to select the most informative samples as the candidate set. Experimental results prove that RLDE can obtain the highest classification accuracy and requires the smallest feature information dimension.

Wang et al. (2019) proposed a semi-supervised learning algorithm based on a deep generative model (Caps-TripleGAN). Generative adversarial network (GANs) has provided a new pathway for sample generation by the use of an adversarial process to perform a semi-supervised learning task in remote sensing classification. Experimental results demonstrate that the reliable generator in TripleGAN can improve the performance of the capsule network. Table 5 summarizes the OA and kappa coefficients of different semi-supervised approaches for remote sensing images.

Active Learning Applications to Remote Sensing Image Classification

Labeling sample is always the first step in a supervised remote sensing image classification task. Unfortunately, the collection of appropriate training sets is not an easy task. To practically deal with this problem, active learning (AL) has been introduced in recent years, with the aim of utilizing the information available from unlabeled data (Tuia et al. 2009). The protocol to label the originally unlabeled data in AL is usually handled by user according to uncertainty, diversity, or/and representative measures (Samat et al. 2015). Specifically, uncertainty measures include margin sampling (MS), multi-class level uncertainty (MCLU) (Demir et al. 2010), entropy query by committee (EQC) (Mamitsuka 1998), and breaking ties (BT) (Luo et al. 2005). And the diversity criterion includes angle-based diversity (ABC), clustering-based diversity (CBD), and enhanced CBD (ECBD) (Demir et al. 2010; Tuia et al. 2011). The representative measures can be found in Huang et al. (2010b) and Samat et al. (2016a, b).

Different from these probability techniques, the possibility approach is an uncertainty analysis tool with imprecise probabilities, commonly used to deal with vagueness and imprecision about information. Therefore, it has great potential for solving vagueness and imprecision issues in querying unlabeled samples in AL.

In fuzzy logic theory, aggregation operators are generally used for aggregating fuzzy sets and fuzzy relations. In a multi-class active learning (MCAL) scenario, if the possibility memberships of unlabeled samples are treated as fuzzy sets, then the aggregation operators in fuzzy logic could be used as the informative

Table 5 The OAs and kappa coefficients obtained by different semi-supervised approaches

Methods	OA	Kappa
SS-LPSVM (Wang et al. 2014)	75.88%	–
S^2SVMSE (Tan et al. 2014)	90.94%	0.878
MLR + KNN + SNI (Tan et al. 2015)	85.47%	0.8038
TT_AL_MSH_MKE (Tan et al. 2016)	85.03%	0.8314
Caps-TripleGAN (Wang et al. 2019) (5% of training samples)	93.58%	0.9047
RLDE_tri_training (Ou et al. 2019)	98.39%	0.978
Dópido's method (Dópido et al. 2013)	84.08%	0.795
Bai's method (Bai et al. 2012) (15 samples/class)	82.85%	0.7806

The default number of initial training samples per class is 10

measurements to evaluate the informativeness of these samples (Wang et al. 2015). Motivated by the previous statements, the fuzzy 2-order ambiguity (F2OA) (Frélicot et al. 2004) and fuzzy C-order ambiguity (FCOA) were proposed to evaluate the classification risk of samples. However, the difference possibility memberships calculation adopted in FCOA can be easily degraded into nonlinear weight function. To solve this issue, Samat et al. proposed a modified FCOA (MFCOA) by following the rule of minimum differences equal to maximum uncertainty in a multi-class scenario and incorporating the normalization of class frequency values to enhance uncertainty measures of unlabeled samples (Samat et al. 2016a, b). Figure 3 depicts the learning curves for all different AL heuristics applied to spectral-spatial features of the ROSIS University dataset. Table 6 reports the average OA and kappa coefficient at the 28th iteration step (to better show the improvements from MCA and MFCOA purpose).

According to the results in Fig. 3, it can be observed that the fundamental possibility approaches CA and FCOA are capable of reaching the upper OA values, with respect to the standard BT, MS, and MCLU. And FCOA shows better learning rates (see the learning curves in green vs. those in magenta in Fig. 3). Furthermore, if we compare the results from the original CA and FCOA approaches, MCA and MFCOA improve with a higher average and lower standard deviation (green vs. cyan, and magenta vs. yellow). For instance, MFCOA reached 96.34% OA at the 28th query with 198 labeled samples for ROSIS Univ. data.

Based on the experiments, it can be stated that:

1. CA and FCOA are capable of reaching the same accuracy of a fully trained SVM with a much smaller training set, and FCOA performs better than CA.
2. With respect to the original CA and FCOA, MCA and MFCOA provide better learning rates and higher classification accuracy values.
3. The performance of CA, MCA, FCOA, and MFCOA are further improved by exploiting diversity criteria.

Ensemble Learning for Improving Unstable Algorithms

Basics of Ensemble Learning

In the pattern recognition field, there is no classifier that can always achieve the best result. However, better performance might be achieved through ensemble learning (EL) when compared to a single classifier. Different from other learning machine techniques, ensemble learning constructs many sub-learners to find practical solutions for a specific problem. The final solution is generated by integrating partial solutions from sub-learners (Kuncheva 2002). In recent years, ensemble learning has been gaining popularity in various fields, especially in the big data era (Al-Jarrah et al. 2015), such as computer vision (Renda et al. 2019), medical and remote sensing applications (Du et al. 2012a, b; Kumar et al. 2016), due to the ability to provide better and stable performance.

With its superiorities of simplicity and effectiveness, ensemble learning has become one of the crucial problem-solving techniques in image and signal processing. Furthermore, the following motivations are also attracted to select ensemble learning (Kuncheva 2002): (1) to avoid determining the initial parameters of each learning machine, (2) to inject the randomness to the learning process and produce the various outputs of each learning machine, which is beneficial for the ensemble, and (3) to use complementary learning machine to improve dynamic adaption and flexibility.

In ensemble learning, an individual learning machine is treated as the base learner. Diversity and accuracy of the base learner are the two essential components in ensemble learning. A powerful ensemble learning system should include the high accuracy of base learners as well as the high diversity within the ensemble. Following the steps of machine learning,

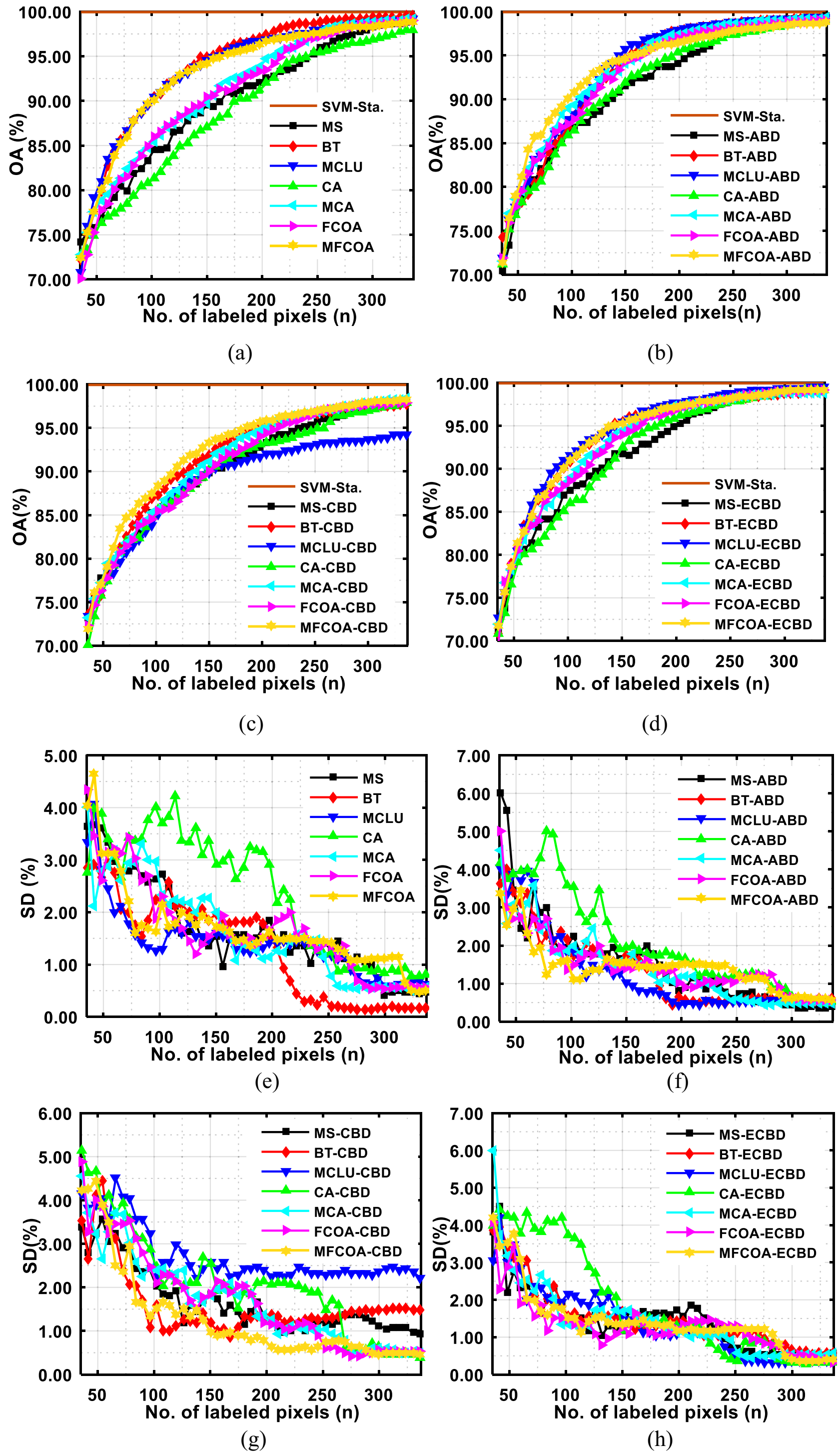


Fig. 3 Learning curves for FMCLU, MS, BT, and MCLU with various diversity criteria (**a**, **e** none; **b**, **f** ABD; **c**, **g** CBD; **d**, **h** ECBD) for ROSIS Pavia Univ. data. Each curve shows the average (**a**–**d**) or standard deviation (SD) (**e**–**h**) values of the OA plotted against the increasing size of the training sets over ten runs of the AL algorithms

diversity could be enforced by the manipulation of training instance, input features, and learning machines.

Combination of Multiple Classifiers for Remote Sensing Classification

Multiple classifier systems (MCSs) have been proved to be an effective solution in improving recognition performance (Su et al. 2014; Fauvel et al. 2013; Du et al. 2012). Specifically, fusion-based and selection-based methods are widely used to integrate the classifiers. For fusion-based method, it can adopt base classifiers in parallel and combine the results to achieve consensus (Woods et al. 1997). To improve the performance of classification, individual error of each classifier is required to form an ensemble system (Tumer and Ghosh 1996), which is difficult to achieve. For selection-based method, it directly

Table 6 Average overall accuracy (OA) and kappa coefficient (κ) values for different AL methods and diversity criteria applied to the ROSIS Univ. dataset (number of labeled samples = 198)

Diversity	None		ABD		CBD		ECBD	
Statistics	OA (%)	*Kappa*	OA (%)	*Kappa*	OA (%)	*Kappa*	OA (%)	*Kappa*
MS	91.99	0.91	94.05	0.93	92.72	0.92	94.96	0.94
BT	96.97	0.97	97.81	0.98	95.16	0.95	97.03	0.97
MCLU	96.71	0.96	97.97	0.98	91.71	0.91	97.70	0.97
CA	91.22	0.90	95.33	0.95	93.05	0.92	95.86	0.95
MCA	93.91	0.93	97.43	0.97	94.79	0.94	97.29	0.97
FCOA	93.37	0.92	97.02	0.97	93.98	0.93	96.75	0.96
MFCOA	96.34	0.96	96.44	0.96	95.61	0.95	97.18	0.97

chooses the classifier with the best performance from an ensemble of classifiers for a given pattern (Giacinto and Roli 1997). Moreover, if one classifier in an ensemble strongly dominates others, selection-based methods have superior performance over fusion-based methods (Kuncheva 2002).

There are usually two types of selection-based methods, which include static classifier selection (SCS) and dynamic classifier selection (DCS). The difference is that the best classifier selected by SCS is for all patterns, while the classifier chosen by DCS only best suits for the single pattern to ensure a reasonable classification result (Ko et al. 2008). Most research tries to utilize spectral information of hyperspectral images for classifier combination (Smits 2002; Didaci et al. 2005). Recently, spatial information (Tarabalka et al. 2010a; Benediktsson et al. 2005, 2007) has attracted more interests. However, spatial and spectral information are just used as data sources of the classifier in most DCS-related work. In practice, pixels are spatially related; in other words, it is highly probable that two adjacent pixels belong to the same class. Therefore, considering both spectral and spatial information can improve hyperspectral imagery classification accuracy significantly.

Su et al. (2014) proposed a novel dynamic classifier ensemble method by combining spectral and volumetric textural features. Spatial and spectral information are used to determine the label when the classified pixels' percentage of unlabeled pixels' neighborhood meets the specific threshold. Pixels where all member classifiers agree are assigned labels directly according to the classification result of each member classifier. For any remained pixels, the proportion of labeled pixels in an $L \times L$ window (e.g., $L = 5$) around the unlabeled pixels is calculated. If the proportion exceeds the threshold (set as 70%) (Riitters et al. 2009), which means the spatial information of the current pixel is strong enough to assign a label to the current pixel. Meanwhile, minimum estimated abundance covariance (MEAC)-based band selection and the volumetric gray-level co-occurrence matrix (VGLCM) model are used to extract spatial and spectral features for dynamic classifier ensembles.

In the experiments, DCS-LCA method is used for comparison, whose goal is to estimate each individual classifier's accuracy in local regions surrounding a test sample and use the decision of the most locally accurate classifier. Local accuracy is estimated according to output classes. DCS involve SVM with Gaussian and RBF kernel is used as the classifiers, also KNN, and Diagquadratic classifiers are employed in the experiments. For texture extraction, the best box size to describe the dataset is $9 \times 9 \times 9$ for VGLCM. In the experiment, the extracted textural features are combined with the original dataset, 5 PCs, and 15 selected bands using the MEAC algorithm. The classification results are reported in Table 7. We can see that the proposed method outperforms other methods.

Rotation Forest and Improvements

Rotation forest (RoF) adopts data transformation and random subsets, which aims at enhancing diversity within the ensemble and the accuracy of the base classifiers simultaneously. The main steps are as follows (Rodriguez et al. 2006):

1. Disjoint subsets are formed by randomly splitting the feature space.
2. Data transformation (e.g., PCA) is utilized to each feature subset with the sub-training set which is bootstrap selected from the original training samples (75% size).
3. A sparse rotation matrix is created by concatenating the coefficients of the principal components in each subset.
4. A single base learner (e.g., decision tree) is utilized for classification based on the new training set reformed by the sparse rotation matrix.
5. The final result will be obtained by combining the output of individual base learners.

To improve the performance of rotation forest, several strategies are involved:

1. Using local data transformation (e.g., locality preserving projections, LPP) instead of PCA (Xia et al. 2014a, 2015a).

Table 7 Classification results using different methods

		SVM	DCS-LCA	The proposed DCS method
All bands	OA	0.9495	0.9442	0.9539
	Kappa	0.9388	0.9322	0.9440
VGLCM 6 + all bands	OA	0.9522	0.9535	0.9608
	Kappa	0.9422	0.9454	0.9524
VGLCM 6 + 5 PCs +5 selected bands	OA	0.9512	0.9524	0.9525
	Kappa	0.9408	0.9411	0.9416
VGLCM 6 + 5 selected bands	OA	0.9503	0.9512	0.9526
	Kappa	0.9397	0.9401	0.9425

2. Using fast and reliable base learner (e.g., support vector machine, random forest), instead of the decision tree (Xia et al. 2015b).
3. Using more informative features like extended morphological attribute profiles (EMAPs) instead of spectral information or regularization methods (e.g., Markov random fields, MRFs) (Xia et al. 2015a, b).

Table 8 presents the classification accuracies of rotation forest and its improved versions in our past work. All the experiments are conducted by using the standard training and testing samples of the University of Pavia ROSIS dataset to make fair comparisons. From Table 8, it can be seen that RoFs with different base learners (e.g., decision tree, random forest, and ELM) have shown better performance than other ensemble classifiers, such as Bagging, AdaBoost, and Random subspace. Since LPP introduces more diversity than PCA into the RoF, thus, RoF-LPP outperforms RoF-PCA. Moreover, with the help of EMAPs and MRFs, RoF-LPP-MRFs, RoRF-EMAPs, and RoELM-EMAPs have shown significant improvements than other classifiers.

Ensemble size and number of features in a subset are the main parameters of RoF, also regarded as indicators of the operating complexity. Figure 4 shows a sensitivity analysis of the two parameters. The accuracies are slightly increased when the ensemble size increases. Here, we suggest adopting the medium size (e.g., 20–40) to make a balance with accuracy and computational complexity. For ROSIS image, the classification performances of RoF are decreased when the number of features in a subset increases. It should be noted that the changing trend of this parameter is diverse in different particular applications.

Ensemble Extreme Learning Machines to Enhance a Weak Classifier

Conventional artificial neural networks (ANNs) are effective nonlinear ML methods with applications to plenty of fields. But any successful applications have to face and tackle the issues of computational inefficiency, network structure complexity, and over-fitting, and plenty of solutions have been proposed (Aguiar et al. 2015; Vardhana et al. 2018). Among those, extreme learning machine (ELM) is one solution for the bottleneck of the learning speed of single-hidden layer feed-forward neural networks (SLFNs), by introducing the Moore-Penrose generalized inverse of matrix technique (Huang et al. 2006). And in contrast with the algorithms like BP neural network and SVM, the performance of ELM on both classification and regression problems is promising. However, ELM also has some drawbacks that the randomness of input weights and bias can cause ill-posed problems, which leads to low performance or even no solution scenario. Thus, ELM can be used as a weak learner in ensemble learning (EL), similar to other weak learners as neural networks and decision tree (Samat et al. 2014; Samat et al. 2015; Du et al. 2014). In our work, ensemble learning is integrated with ELM and two implementations of ensemble extreme learning machines (E2MLs) based on Bagging and Boosting, BagELMs and BoostELMs, are proposed for hyperspectral image classification (Samat et al. 2014).

In Fig. 5, we present the computational costs of the considered algorithms on the original and the EMP features of ROSIS Pavia University and Salinas. It can be easily noticeable that the ELM algorithms are much faster than SVM for the step of parameter search. However, in general, ELM methods consumed more time for training and classification. Nevertheless, for the whole process, ELM methods are faster than RBF kernel-based SVM.

In Table 9, classification accuracy and computational measures are present for considered algorithms. From this table, one can easily see that SVM shows higher OA values than ELM, but close to BagELMs and BoostELMs in most cases of using spectral or using stacked features of spectral with EMPs. In the meantime, SVM takes huge time for searching the optimal parameters in a grid search-based criterion. Additionally, the classification accuracies of ELM, BagELMs, and BoostELMs are enhanced as well by including the spatial information. The best performance is reached by BagELMs or BoostELMs using the first 10 principal components or using the first 10 PCs with EMP features.

Table 8 Classification accuracies of rotation forest and its improvements

Methods	OA	AA	Kappa
Bagging	65.78	77.42	58.71
AdaBoost	67.81	78.84	61.36
Random subspace	70.44	81.78	63.90
Random forest	71.37	81.93	64.79
ELM	74.56	84.64	68.75
RoF-PCA	82.66	87.69	78.09
RoF-LPP	84.76	88.64	79.35
RoRF	79.04	87.27	73.98
RoELM	79.44	87.11	74.25
RoF-LPP-MRFs	92.15	93.06	89.27
RoRF-EMAPs	96.47	95.84	95.34
RoELM-EMAPs	98.69	98.92	98.25

Two basic observations can be summarized: (1) ELM-based methods obtain better or comparable accuracies in comparison with SVM and (2) BoostELM and BagELMs outperform the original ELM, showing that ensemble strategies are useful for improving weak learner like ELM.

Ensemble Learning for Spatial Interpolation

Spatial interpolation is one of the most important methods to show the spatial differentiation pattern of soil properties. Currently, the spatial interpolation methods of soil properties are mainly derived from the models of regression, surface simulation, and probability theory, and the interpolation models require many simplifications and assumptions on the data (Yue and Wang 2010). However, soil properties have significant spatial differentiation in complex geomorphological areas. Therefore, the assumptions of the existing interpolation model cannot be satisfied, and the shortage of the single interpolation model limits the prediction accuracy (Goovaerts 2011). In order to solve the discontinuity and spatial variation of soil properties in complex terrain areas, the ensemble learning-based spatial interpolation method is constructed that can adaptively divide the interpolation surface to screen the appropriate basic interpolation model (Liu et al. 2016, 2017). In this section, the basic interpolation model of the screening is optimized, and the multi-model ensemble interpolation method (SP-EL, the soil property surface modeling based on ensemble learning) is established and coordinated to simulate the soil properties with high precision (Liu et al. 2018).

Interpolation Surface Partitioning

A series of soil property interpolation surfaces are generated using a built-on interpolation model, such as Radial Basis Function, Cokriging, and empirical Bayesian. The interpolation surface is scanned with a scan line algorithm as a learner of the ensemble learning process (Fig. 6), and the prediction error of the sample point of the soil property is calculated. Each interpolated surface is adaptive to be partitions (Fig. 6), thereby obtaining the applicable spatial range for each interpolation model (Figs. 6, 7, and 8).

Interpolation Surface Combination

As can be seen in Fig. 9, the spatial range of soil property in ordinary kriging (OK), regression kriging (RK), and

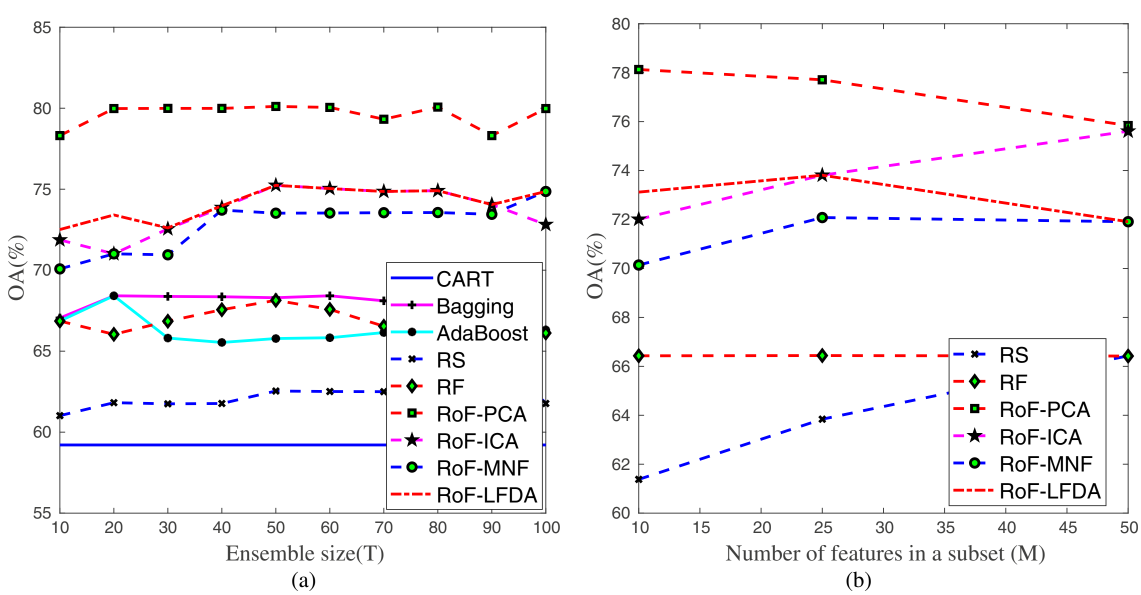
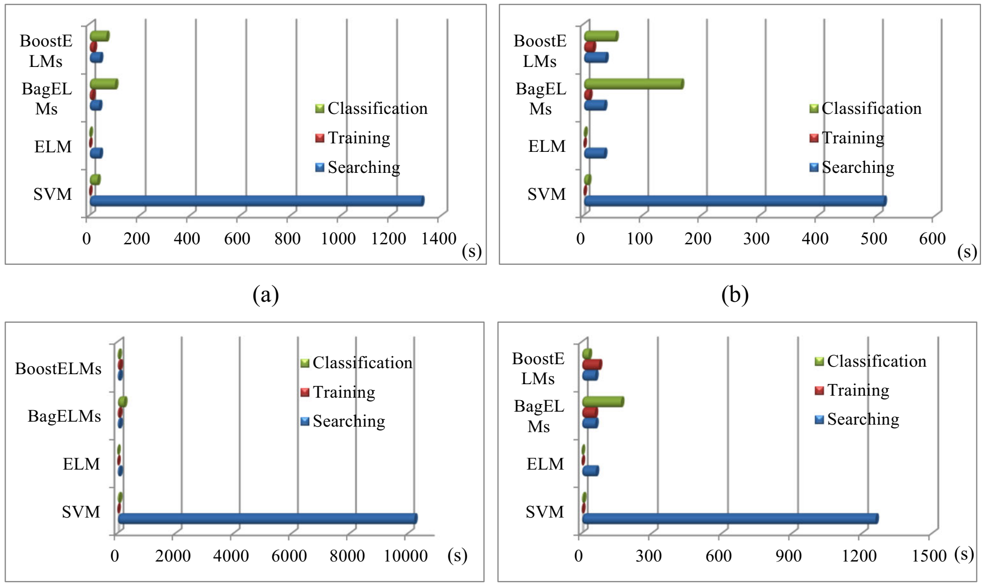


Fig. 4 Parameter analysis of **a** ensemble size and **b** number of features in a subset



(c) (d)

Fig. 5 Computational costs of ELM, BagELMs, BoostELMs, and SVM on original (**a**, **c**) and EMPs (**b**, **d**) features of ROSIS Pavia University (**a**, **b**) and Salinas (**c**, **d**) datasets (horizontal axis represents time consumption in seconds)

Bayesian kriging (BK) is smaller than the measured value, and OK is the smallest. The interpolation maps of RK, BK, and OK showed varying of weak "bull's-eye" effects. What's more, RK and BK interpolation can better depict the spatial variation of soil properties, but shows the local changes of soil property content with difficulty. The inverse distance weighting (IDW) interpolation presents a strong "bulls-eye" effect with the worst result. SP-EL best describes the spatial pattern change of soil property and produces a more appropriate range of interpolation. SP-EL can show the spatial pattern change of soil characteristics in more detail. In particular, it can precisely show the mutation of soil property content.

Deep Learning for Scene Classification and Urban Structural Type Recognition

Basic Scene Understanding Algorithms in Deep Learning

With the advancement of remote sensing observation technology, high-resolution sensors can collect images with a spatial resolution of finer than 1 m. These high-resolution images with abundant spatial and structural patterns provide the possibility of understanding high-level land cover and land use information or target of interest for surface monitoring and management. To

Table 9 Classification accuracies (overall accuracy) and computational time (s) of using ROSIS data

Feature		All spectral bands				All spectral bands with EMPs			
Classifier		SVM	ELM	BagELMs	BoostELMs	SVM	ELM	BagELMs	BoostELMs
Overall accuracy (%)		*80.3*	74.98	75.39	79.11	83.7	81.79	83.98	*84.66*
Kappa statistics		0.75	0.69	0.69	0.74	0.70	0.77	0.80	0.80
Time	Searching	1321	42.39	39.67	43.28	3283	53.93	50.17	54.15
	Training	0.532	0.37	12.68	18.58	0.78	0.07	3.82	19.4
	Classification	33.17	2.32	103.07	68.74	62.14	1.32	174.52	92.11
Feature		First 10 PCs				First 10 PCs with EMPs			
Classifier		SVM	ELM	BagELMs	BoostELMs	SVM	ELM	BagELMs	BoostELMs
Overall accuracy (%)		82.32	82.16	*83.29*	80.79	93.66	92.31	93.68	*94.3*
Kappa statistics		0.78	0.77	0.77	0.76	0.92	0.90	0.92	0.93
Time	Searching	511.9	35.21	35.28	37.52	600.05	36.77	36.87	37.4
	Training	0.63	0.34	9.92	16.09	0.19	0.22	15.54	16.94
	Classification	8.1	1.81	166.37	54.69	6.45	1.64	203.96	57.84

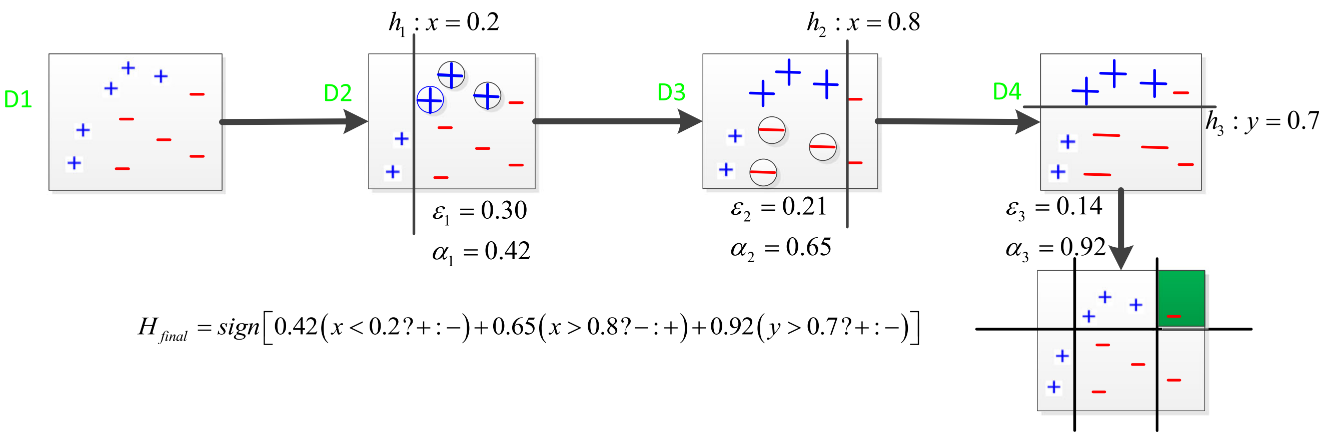


Fig. 6 Interpolation surface partition process

extract this valuable information from remote sensing data, one of the fundamental techniques is scene classification (Wang et al. 2019). Scene classification, as a typical issue of computer vision, is also considered as a challenging task in remote sensing domain because of some special properties of remote sensing scene images, such as various spectral and spatial resolution, complex composition in land cover categories, and uncertain size of the scene unit. However, the biggest challenge is to construct an effective feature representation based on original image to describe the semantic information.

Over the years, many studies have been devoted to developing robust feature learning and representation methods. In the early years, some feature extraction methods based on low-level visual features (e.g., color, texture, and shape features) or manual designed feature descriptors (e.g., local and scale invariant feature transform) have been conducted to finish the remote sensing scene classification (Yang and Newsam 2010). In the feature representation stage, some advanced feature coding methods, including bag-of-words (BOW) and Fisher kernel vector, and topic models like probabilistic latent semantic analysis (PLSA) have been employed (Cheng et al. 2013; Zhao et al. 2016). Moreover, more similar studies focused on improving performance in consideration of the spatial arrangement of scene images. In this scheme, spatial pyramid match kernel and its improvements have been proposed (Yang and Newsam 2010; Lazebnik et al. 2006). These methods have acquired satisfactory results for some simple test datasets, which always include few categories and scene images of fixed size with uniform structures, spatial textures, or discriminative color features of different scene objects. However, it is still a challenge to deal with the practical problem using traditional algorithms.

Recently, deep learning-based methods have achieved impressive results in scene image classification and outperformed traditional methods. Inspired by the information abstraction process of the human brain, deep learning utilizes different functional layers to construct a depth architecture simulating the brain cognitive process. Initial deep learning research focused on feature learning with an unsupervised manner that pays more attention to feature reconstruction than category discrimination with a

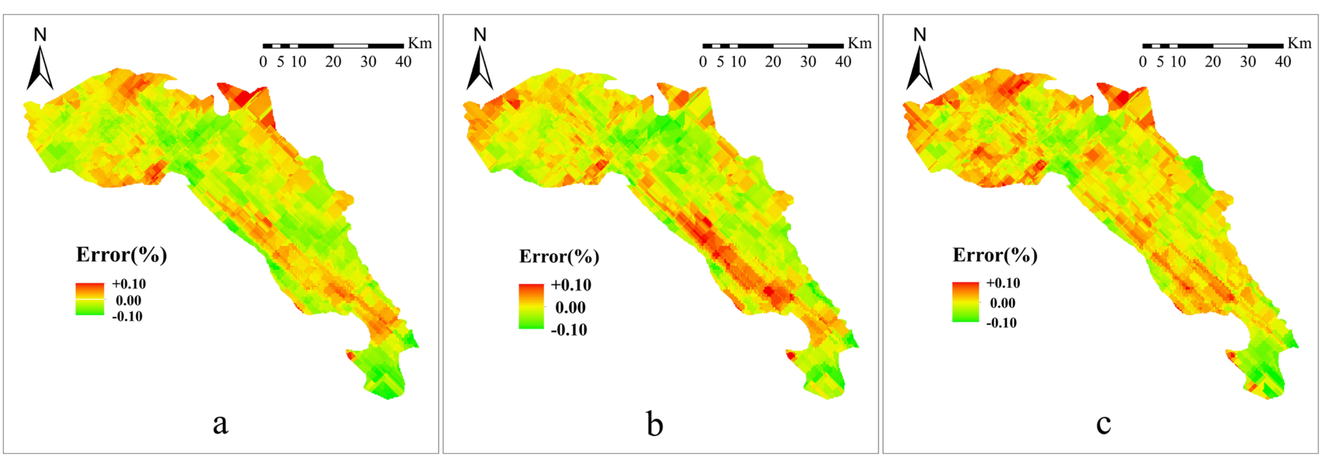


Fig. 7 Error surface of basic interpolation model: **a** surface a; **b** surface b; **c** surface c

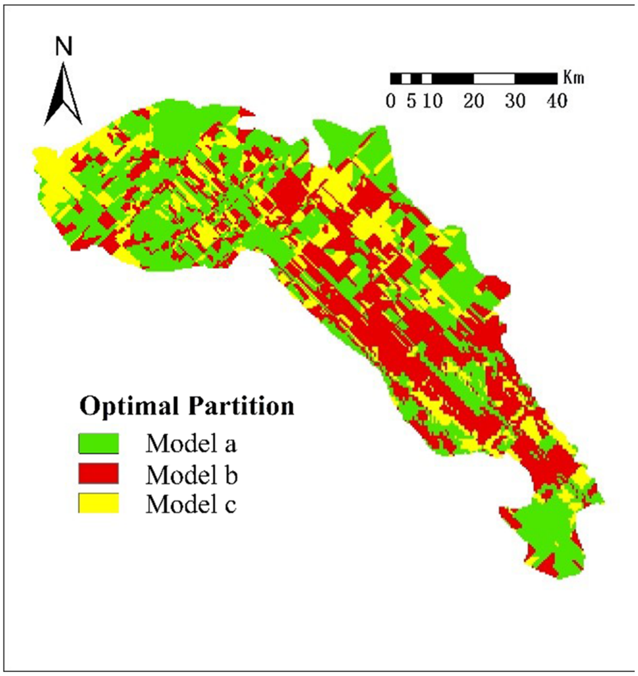


Fig. 8 Optimization distribution

shallow network, such as restricted Boltzmann machines (RBMs) and autoencoder (AE). However, the breakthrough is the proposal of deep belief nets (DBNs) (Hinton and Salakhutdinov 2006), which is composed of multiple RBM models and a logistic regression classification layer. In the training stage, DBNs train one RBM at a time in an unsupervised manner and then fine-tuning the whole networks in a supervised manner. This research shows the importance of the unsupervised feature learning process in developing a successful deep architecture and overcoming the overfitting problem.

Motivated by this work, a number of unsupervised feature learning algorithms are proposed, such as sparse coding and sparse AE (Cheriyadat 2013; Li et al. 2017). Significantly, these unsupervised feature learning methods attempt to exploit essential feature representations of raw data by using a shallow network model. Moreover, it is easily extended to a supervised algorithm by stacking a classifier layer, or used to construct a deep architecture by concatenating multiple unsupervised models and a classifier. Besides, as the most representative supervised deep learning model, deep convolutional neural network (CNN) also achieves great success in scene classification. The deep structure of CNN has the ability to learn hierarchical level abstraction of input data by adopting the structure of the multi-layer networks. Furthermore, CNN-based algorithms integrate feature learning and category

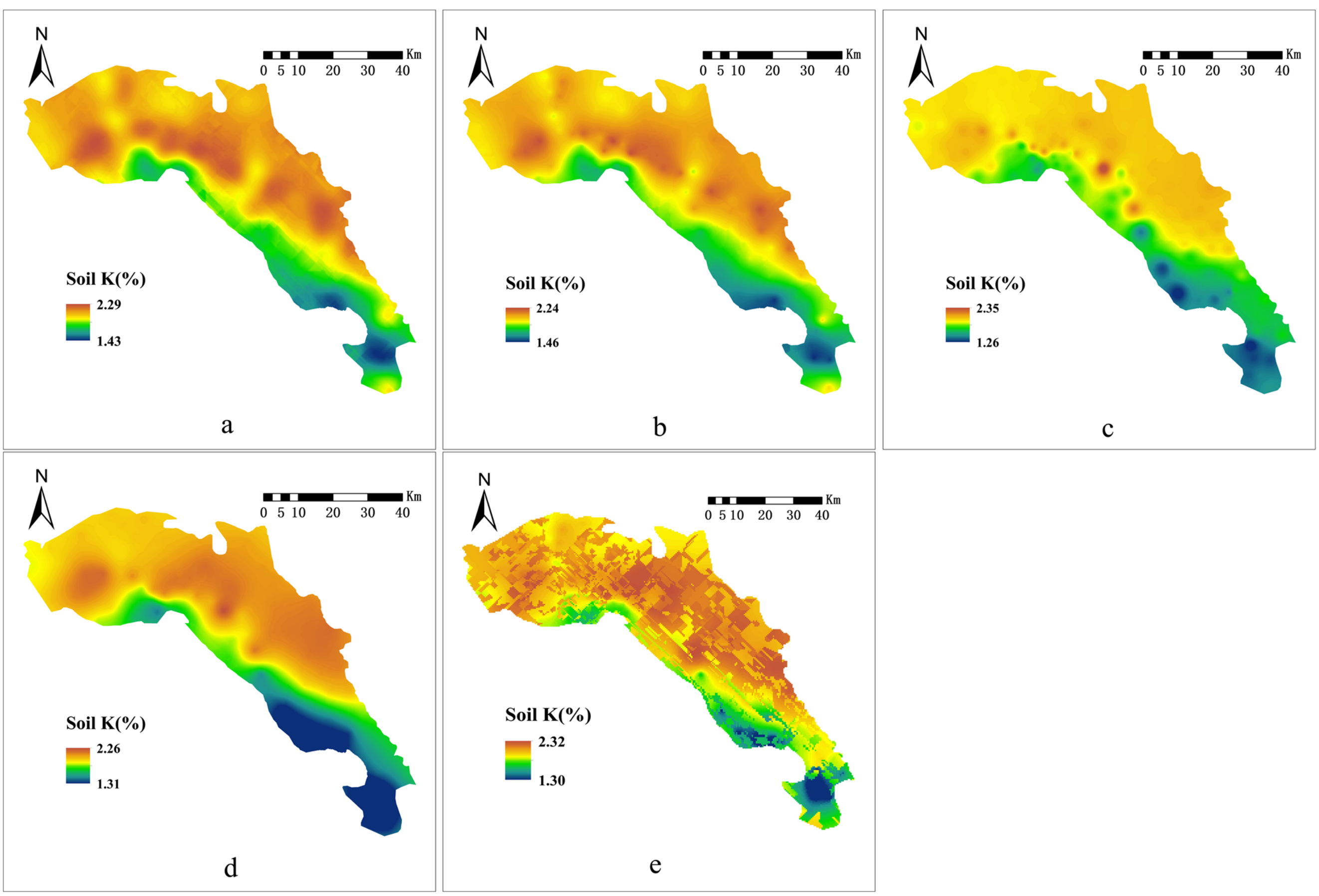


Fig. 9 Comparison of soil property content interpolation map with different methods. **a** RK. **b** BK. **c** IDW. **d** OK. **e** SP-EL

discrimination into one framework. In this review, we provide a brief introduction to the unsupervised and supervised methods and its applications in remote sensing scene classification. More detailed information on deep learning theory in the computer vision and machine learning community can be found in Krizhevsky et al. (2012).

Unsupervised Feature Learning-Based Methods

There are diverse unsupervised methods for feature learning and scene image classification. Among these methods, sparse coding is a typical unsupervised method that efficiently represents data by learning sets of over-complete bases (Cheriyadat 2013). Use an over-complete base to represent input vectors, sparse coding can effectively capture structures and patterns inherent in the input data. Sparse coding has been widely exploited to learn feature representation of images, multiple features, and appearance descriptors such as SIFT features (Cheriyadat 2013). In addition, RBM as a layer-wise training model in the construction of a DBN is another kind of unsupervised method. It is the improved version of Boltzmann machines (BMs) (Hinton 2002). BMs are a particular form of log-linear Markov random field with visible units and hidden units. Once an RBM model is learned, input data is encoded by the weights matrix in the hidden layer to represent the raw data.

In remote sensing scene classification, RBM is also employed to learn internal feature representation for scene understanding (Zou et al. 2015). Besides, the sparse AE method also attracted much attention in the remote sensing community. Sparse AE is a symmetrical neural network with only one hidden layer. It trains an AE with a given number of hidden nodes using backing propagation to minimize squared reconstruction error between input values of the encoding layer and output values of the decoding layer. A sparse AE model includes an encoder and a decoder, in which input data can be encoded by the encoder to carry out the feature representation. Furthermore, in order to force the hidden layer to discover more robust features and prevent it from simply learning the identity, the denoising AE is proposed. The denoising AE is a stochastic version of the AE model (Vincent et al. 2008). In practice, the stochastic corruption process randomly sets some of the inputs (as many as half of them) to zero and then fed them into a sparse AE to train a denoising AE model. In another study, a new unsupervised method, triangle encoding, is proposed. When the number of hidden nodes is set to a large value, this method can obtain the best performance (Coates et al. 2011).

Generally, the scheme of unsupervised feature learning method for remote sensing scene classification consists of five main steps: (1) image patch or local feature extraction; (2) unsupervised feature learning model training; (3) feature mapping of scene images using learnt unsupervised feature learning model through convolution operation; (4) feature coding or middle-level feature representation; and (5) scene classification. This scheme can be summarized as in Fig. 10. In the first step, image patches are randomly extracted from all scene images for training an unsupervised feature learning model. It is worth noting that image patches are not the only option as the input to train the model, other features with the possibility to describe the characteristics of scene images can be used as well. Then, a selected unsupervised feature learning method is employed to train a required feature learning model based on the extracted features. Thirdly, an input scene image can be mapped into a new feature space with the learned model through convolution operation. Up to now, the raw scene images are represented by the new feature maps, which are still three-dimensional data. Hence, we must use a feature coding method, such as BOW, Fisher kernel (Perronnin and Dance 2007), or pooling method, to encode the three-dimensional data into vector representation form. Finally, the vector representation of scene images as input data is used to train a classifier to predict the categories of scene images.

In this section, we present some experimental results of remote sensing scene classification based on the unsupervised feature learning method. In this experiment, image patches are extracted to train the feature learning model, and sparse AE is employed as the unsupervised feature learning method. At the feature coding stage, some commonly used feature coding methods like max-pool, Fisher kernel (FK), and BOW are investigated. Furthermore, we also propose a feature coding method based on global feature descriptors (Li et al. 2017). The widely used dataset UC Merced (UCM) dataset is used to evaluate the unsupervised method (Yang and Newsam 2010). It consists of 21 land use categories, and each category contains 100 images of size 256 × 256 × 3. Following the common experimental setup in related studies, all samples of input dataset are randomly divided into five equal parts, and four parts are used as training samples and the rest as a testing set every time. We train a sparse AE model with 400 hidden units for feature learning. For comparisons with different feature coding methods, the proposed feature coding method global feature coding (GFC) (Li et al. 2017), maximum pooling (max-pool), BOW, and FK are employed to build feature representation for classification. As shown in Table 10, among the four feature coding methods, GCF achieves the highest accuracy. These experimental results also indicate that the feature coding process is an important factor for unsupervised feature learning-based scene classification.

Supervised Feature Learning-Based Methods

Building a deep supervised neural network is always a pursuit in machine learning study. Generally, the earliest back-propagation neural networks are the antecedents of deep supervised neural networks. However, it is widely believed that

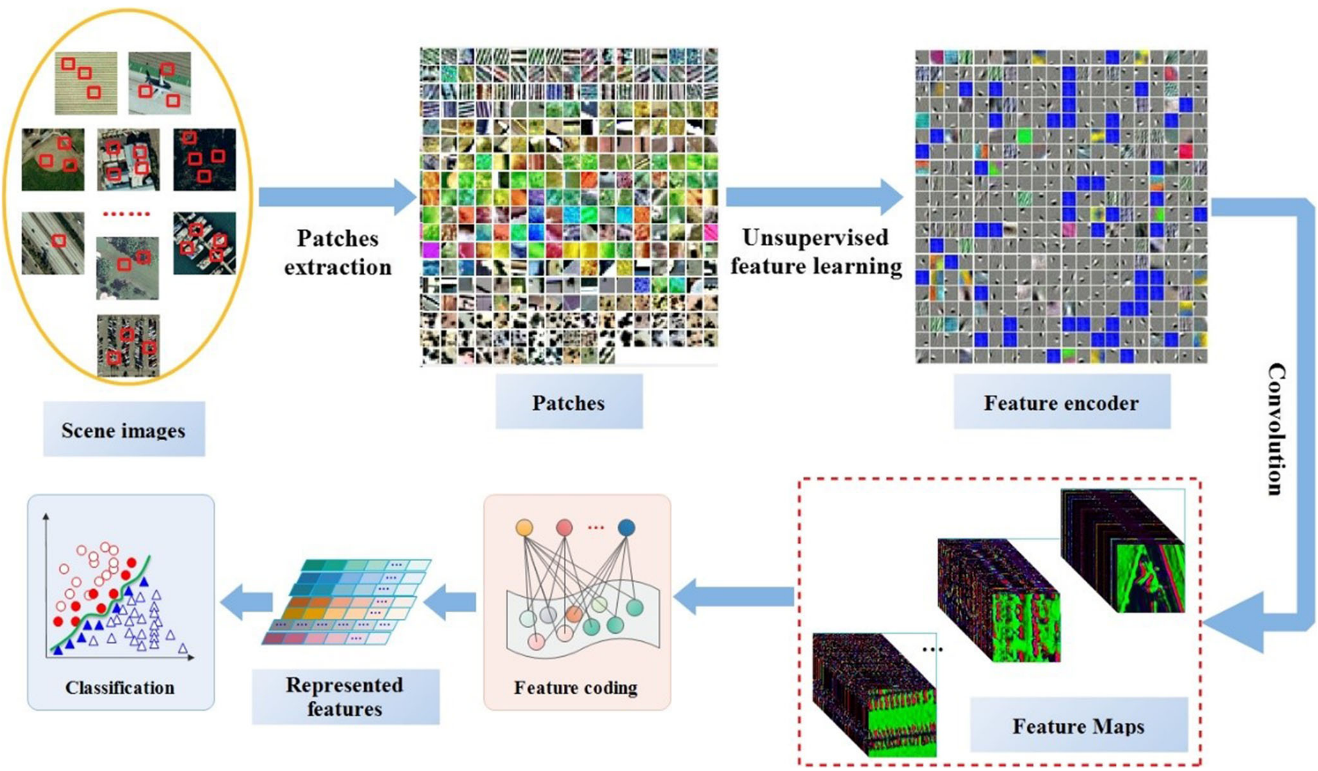


Fig. 10 The overall flow chart of the unsupervised feature learning-based scene classification

training a supervised neural network with very deep architectures is difficult before the appearance of DBNs. Recently, it has become an easy task in constructing and training a deep network due to the advancement in hardware and software technology, and big data acquisition ability. There are two patterns to build a deep network. The first one is stacking series of unsupervised networks and a classification layer to complete a supervised manner, with DBNs as the typical example of this pattern is. Actually, other unsupervised feature learning like sparse AE is also employed to build a deep supervised network (Abdi et al. 2017). Besides, another pattern of deep learning networks is deep convolutional neural network (CNN). CNN is the most popular deep learning structure, which is widely used for learning visual features in several distinct tasks, such as remote sensing scene classification (Penatti et al. 2015). In contrast with an unsupervised feature learning-based scene classification scheme, CNN integrates feature learning and classification processes into one framework. Moreover, a multi-layer CNN model can extract different levels of information, ranging from low-level information in the first layer to high-level information in the final layer.

Table 10 Classification performances' comparison using different feature coding methods on UCM dataset

Method	Classification accuracy (%)
max-pool	62.52±2.47
Fisher kernel	79.05±2.01
Bag-of-words	81.19±1.00
Global feature coding	89.62±1.67

Recently, many CNN-based algorithms are proposed for remotely sensed scene classification and achieve better performance than the traditional methods (Xia et al. 2015; Nogueira et al. 2017). Some studies have tried to design a reasonable CNN model to solve remote sensing scene classification (Zhang et al. 2015). A general scheme of scene classification based on CNN models is displayed as Fig. 11. Although the CNN shows an impressive ability to distortion and rotation for remote sensing scene classification, some challenges have been exposed in the application, such as various scales, and complex spatial assignment of remote sensing scene images. Furthermore, it is almost impossible to train a very deep CNN model for many applications with only a few hundreds or thousands of samples. To overcome this problem encountered when using deep CNN models, many studies use transfer learning to address this issue (Nogueira et al. 2017). To this end, the pre-trained CNN models trained on large datasets are employed as feature extractors in remote sensing scene images. In this framework, the pre-trained CNN is only used as a feature extractor and is not trained on the target data, so it can be regarded as an unsupervised learning feature learning process that can prevent overfitting problem. In addition, multi-scales scene images of various sizes can be used as input of the pre-trained CNN model to alleviate the scale problem.

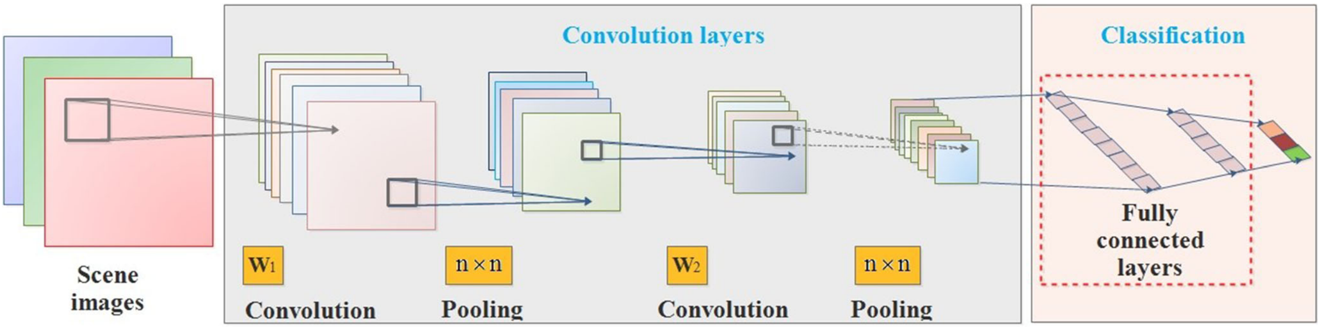
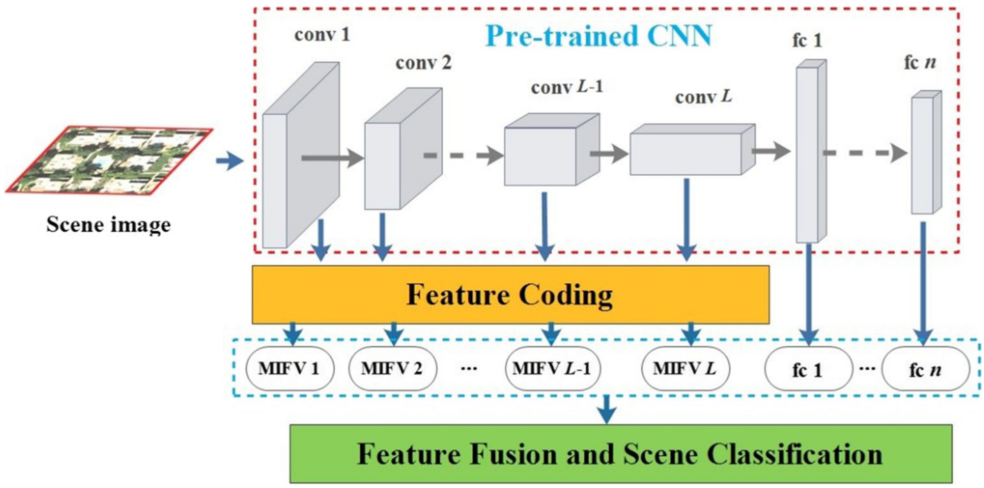


Fig. 11 A generally CNN structure for scene classification

Fig. 12 The flowchart of fusing multi-layer features for scene classification. Conv1-L represents the convolutional layers; MIFV1-L denotes the coded features of different convolutional features; fc1-n represents the fully connected layers



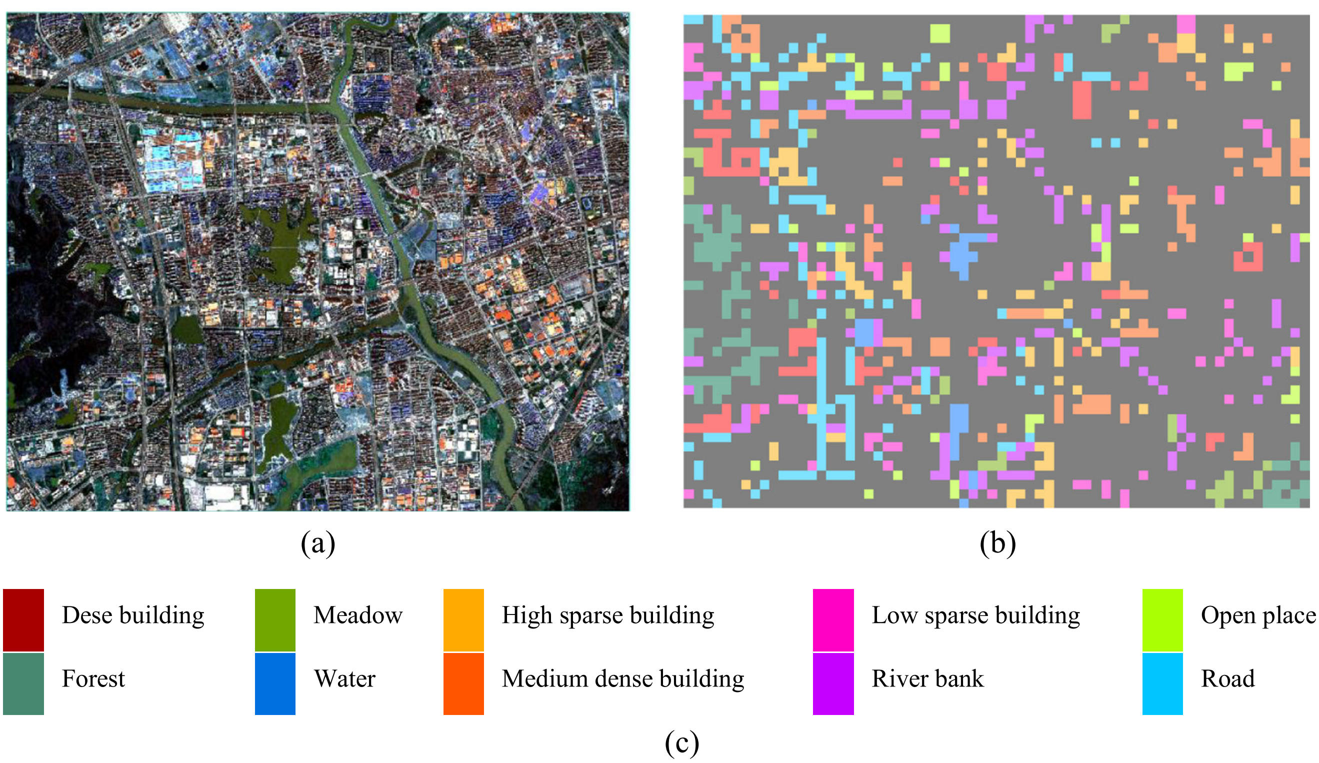


Fig. 13 **a** The SPOT 7 image of being annotated. **b** The image ground truth. **c** Ten urban structural types

Table 11 Detailed information of ground truth data for urban structural types

Class name	Samples	Class name	Samples
Dese building	99	Meadow	37
Medium dense building	106	Water	33
High sparse building	102	Road	128
Low sparse building	75	River bank	146
Open space	52	Total	859
Forest	81		

Motivated by these advantages, some studies also make efforts to exploit the benefits of using multiple layers and pre-trained CNN models (more than three) in remote sensing scene classification (Hu et al. 2015; Li et al. 2017). In the future, more advanced CNN models or improved deep learning method will be employed in remote sensing scene classification domains.

Then, we provide some examples of remote sensing scene classification based on CNN methods. In these experiments, we use the pre-trained CNN models as feature extractors and focus on investigating the performances of different CNN models. Furthermore, some improvements in taking advantage of multi-layer features and CNN models are proposed for remote sensing scene classification. Briefly, a multi-layer feature fusion framework is firstly proposed to integrate multiple level features extracted by a pre-trained CNN model to improve the performance of scene classification. In addition, we also pay attention to exploit the benefits of different CNN architectures and extend the multi-layer feature fusion framework to multiple model fusion. The basic flow of fusing multi-layer features and multiple models are shown in Fig. 12. In this framework, the convolutional features, including lower abstractions of scene image, are encoded by feature coding methods into a vector representation, then stacked with fully connected features for scene classification. With multiple pre-trained CNN models, we also can extract feature representation from every model and stack series features from various models to construct feature representation for scene images. It is important to note that encoded features of convolutional features or fully connected features are always high-dimensional features, so it is a challenge to fuse those high-dimensional features. In our studies, some subspace feature learning methods are proved to be effective in processing high-dimensional features. Furthermore, a supervised subspace learning method can decrease the diversity of multiple models. Therefore, a fusion strategy of supervised and unsupervised subspace learning methods is a more reasonable choice. The more detailed information on feature coding methods and feature fusion strategies can be found in Li et al. (2016, 2017) and Du et al. (2019).

In this experiment, the UCM dataset is used as the experimental data as well. Besides, we also take a test on a large satellite image acquired from Nanjing, China, by SPOT 7 sensor. The spatial resolution of this data is 1.5 m for the panchromatic band and 6 m for multi-spectral bands. They are fused by the Gram-Schmidt pan-sharpening method and reached 1.5 m spatial resolution of the multi-spectral image. The red, green, and blue channels are selected in scene classification. The large image to be annotated is of 5200×6600 in size, as shown in Fig. 13. This experiment is designed to challenge the urban structural type recognition by using the scene classification method. Therefore, ten urban structural types are annotated, including dense building, meadow, high sparse building, low sparse building, open space, forest, water, medium dense building, river bank, and road. We manually label part of the image to obtain series scene images of urban structural type in which each scene image is of the size of 100×100. In the classification stage, 50% of labeled samples are randomly selected as the training set, and the remaining scene images are used for testing. The detailed information of the ground truth data is shown in Table 11.

For the UCM dataset, seven pre-trained CNN models are employed in this experiment. We investigate performances using different layer features. According to the results, the convolution features encoded by feature coding methods achieve better performances than fully connected features. In addition, the best accuracy is always obtained when multi-layer features are fused for classification. Besides, fine-tuning the pre-trained CNN models with the training samples of task maybe improve the performances of CNN models, but cannot outperform the fusion method. As shown in Table 12,

Table 12 Performance comparison of fusing multi-layer features and multiple CNN models for UCM dataset

Pre-trained CNN models	Accuracy (%)	Pre-trained CNN models	Accuracy (%)
AlexNet	98.10 ± 0.48	VGG-F	97.95 ± 0.55
CaffeNet	98.33 ± 0.45	VGG-VD16	98.57 ± 0.24
VGG-M	98.29 ± 0.68	VGG-VD19	97.71 ± 0.69
VGG-S	98.24 ± 0.69	Multi-CNNs	99.05 ± 0.48

Table 13 Performance comparison of fusing multi-layer features and multiple CNN models for SPOT 7 image

Pre-trained CNN models	Accuracy (%)	Pre-trained CNN models	Accuracy (%)
AlexNet	87.04 ± 1.50	VGG-F	87.59 ± 1.32
CaffeNet	88.38 ± 1.13	VGG-VD16	90.37 ± 0.67
VGG-M	87.87 ± 1.02	VGG-VD19	88.75 ± 1.04
VGG-S	87.08 ± 0.60	*Multi-CNNs*	91.11 ± 1.49

fusing multiple CNN models also can improve the performance of remote sensing scene classification.

The performances of urban structural type recognition using scene classification method are shown in Table 13. We compared the classification accuracies of fusing multi-layer features and multiple models by using seven pre-trained CNN models. The results confirm that using CNN-based deep learning method is an efficient way to complete urban structural type recognition.

Conclusions and Prospects

In this paper, we conducted a systematic review in the applications of four advanced machine learning methods for classification and regression-related spatial data handling. Those machine learning-based methods consist of applying kernel learning to address nonlinear and high-dimensional classification and regression problems, adopting semi-supervised and active learning to handle small training sample size challenges, using ensemble learning to combine the advantages of different learners and improve learning robustness, and exploiting deep learning to extract the hidden high-level features. In addition to summarize the advances in these fields, some representative examples completed by the authors are used to demonstrate the merits of advanced machine learning methods for spatial data handling. Based on the technical reviews and experimental analyses, it can be concluded that machine learning methods are suitable to overcome the challenges in spatial data handling and improve the performance of classification, regression, and inversion problems, and will play more and more important roles in the future.

However, it should be noted that both machine learning and spatial data handling methods cover quite a broad range, and only a few of them are reviewed in this paper. For other spatial data handling techniques including point, vector, and map data in GIS, some research has demonstrated the great potential of advanced machine learning methods, for example, deep learning for semantic address matching (Lin et al. 2020), deep neural network for personalized POI recommendation in location-based social networks (Ding and Chen 2018), machine learning models for estimation of land consumption rates (Hagenauer et al. 2019), machine learning for automated generalization of buildings(Steiniger et al. 2010), and hybrid ensemble learning method for tourist route recommendations (Wan et al. 2018).

Other directions in machine learning for spatial data handling should be further explored in the future: (1) embedding the spatial laws and geographical knowledge in machine learning; (2) improving the efficiency of learning and computing by parallel computing or cloud computing platform, for example, Google Earth Engine (GEE); and (3) fusing the advantages of multi-source data and different learning algorithms under open and complex geographical environment.

Funding Information This study is funded by the National Natural Science Foundation of China (Grant No. 41631176).

Compliance with Ethical Standards

The authors confirm the paper complies with ethical standards.

Conflict of Interest The authors declare that there is no conflict of interest.

Ethical Approval This paper focuses on spatial data handling methods and meets the requirements of ethical approval.

Informed Consent The authors confirm informed consent to this submission.

References

Abdi G, Samadzadegan F, Reinartz P (2017) Spectral–spatial feature learning for hyperspectral imagery classification using deep stacked sparse autoencoder. J Appl Remote Sens 11(4):042604

Aguiar LM, Pereira B, David M, Díaz F, Lauret P (2015) Use of satellite data to improve solar radiation forecasting with Bayesian artificial neural networks. Sol Energy 122:1309–1324

Al-Jarrah OY, Yoo PD, Muhaidat S, Karagiannidis GK, Taha K (2015) Efficient machine learning for big data: a review. Big Data Research 2(3):87–93

Ancona N, Maglietta R, Stella E (2006) Data representations and generalization error in kernel based learning machines. Pattern Recogn 39(9):1588–1603

Bai J, Xiang S, Pan C (2012) A graph-based classification method for hyperspectral images. IEEE Trans Geosci Remote Sens 51(2):803–817

Benediktsson JA, Palmason JA, Sveinsson JR (2005) Classification of hyperspectral data from urban areas based on extended morphological profiles. IEEE Trans Geosci Remote Sens 43(3):480–491

Benediktsson JA, Chanussot J, Fauvel M (2007) Multiple classifier systems in remote sensing: from basics to recent developments. In International workshop on multiple classifier systems. Springer, Berlin, Heidelberg, pp 501-512

Bishop CM (2006) Pattern recognition and machine learning. Springer

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory, ACM, pp 92–100

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152). ACM

Camps-Valls G, Gomez-Chova L, Muñoz-Marí J, Vila-Francés J, Calpe-Maravilla J (2006) Composite kernels for hyperspectral image classification. IEEE Geosci Remote Sens Lett 3(1):93–97

Chen Y, Wang G, Dong S (2003) Learning with progressive transductive support vector machine. Pattern Recogn Lett 24(12):1845–1855

Chen Y, Nasrabadi NM, Tran TD (2011) Hyperspectral image classification using dictionary-based sparse representation. IEEE Trans Geosci Remote Sens 49(10):3973–3985

Cheng G, Guo L, Zhao T, Han J, Li H, Fang J (2013) Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. Int J Remote Sens 34(1):45–59

Cheriyadat AM (2013) Unsupervised feature learning for aerial scene classification. IEEE Trans Geosci Remote Sens 52(1):439–451

Chi M, Feng R, Bruzzone L (2008) Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem. Adv Space Res 41(11):1793–1799

Coates A, Ng A, Lee H (2011) An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics, pp 215-223

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3): 273–297

Dalla Mura M, Benediktsson JA, Waske B, Bruzzone L (2010) Morphological attribute profiles for the analysis of very high resolution images. IEEE Trans Geosci Remote Sens 48(10):3747–3762

Demir B, Persello C, Bruzzone L (2011) Batch-mode active-learning methods for the interactive classification of remote sensing images. IEEE Trans Geosci Remote Sens 49(3):1014–1031

Didaci L, Giacinto G, Roli F, Marcialis GL (2005) A study on the performances of dynamic classifier selection based on local accuracy estimation. Pattern Recogn 38(11):2188–2191

Ding R, Chen Z (2018) RecNet: a deep neural network for personalized POI recommendation in location-based social networks. Int J Geogr Inf Sci 32(8):1631–1648

Dópido I, Li J, Plaza A, Bioucas-Dias JM (2012) A new semi-supervised approach for hyperspectral image classification with different active learning strategies. In: 2012 4th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), IEEE, pp 1–4

Dópido I, Li J, Marpu PR, Plaza A, Dias JMB, Benediktsson JA (2013) Semisupervised self-learning for hyperspectral image classification. IEEE Trans Geosci Remote Sens 51(7):4032–4044

Du P, Tan K, Xing X (2012a) A novel binary tree support vector machine for hyperspectral remote sensing image classification. Opt Commun 285(13–14):3054–3060

Du P, Xia J, Zhang W, Tan K, Liu Y, Liu S (2012b) Multiple classifier system for remote sensing image classification: a review. Sensors 12(4):4764–4792

Du P, Samat A, Gamba P, Xie X (2014) Polarimetric SAR image classification by boosted multiple-kernel extreme learning machines with polarimetric and spatial features. Int J Remote Sens 35(23):7978–7990

Du P, Li E, Xia J, Samat A, Bai X (2019) Feature and model level fusion of pretrained CNN for remote sensing scene classification. IEEE J Sel Top Appl Earth Obs Remote Sens 12(8):2600–2611

Ethem A (2010) Introduction to machine learning (second edition). The MIT Press, Cambridge, Massachusetts

Fan RE, Chen PH, Lin CJ (2005) Working set selection using second order information for training support vector machines. J Mach Learn Res 6(Dec):1889–1918

Fang L, Li S, Duan W, Ren J, Benediktsson JA (2015) Classification of hyperspectral images by exploiting spectral–spatial information of superpixel via multiple kernels. IEEE Trans Geosci Remote Sens 53(12):6663–6674

Fauvel M, Benediktsson JA, Chanussot J, Sveinsson JR (2008) Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. IEEE Trans Geosci Remote Sens 46(11): 3804–3814

Fauvel M, Tarabalka Y, Benediktsson JA, Chanussot J, Tilton JC (2012) Advances in spectral-spatial classification of hyperspectral images. Proc IEEE 101(3):652–675

Fotheringham A, Charlton M, Brundson C (2015) Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. Environ Plan 30:1905–1927

Frélicot C, Mascarilla L, Fruchard A (2004) An ambiguity measure for pattern recognition problems using triangular-norms combination. WSEAS Trans Syst 8(3):2710–2715

Frohlich H, Chapelle O, Scholkopf B (2003) Feature selection for support vector machines by means of genetic algorithm. In: Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence,pp 142–148

Ghamisi P, Souza R, Benediktsson JA, Zhu XX, Rittner L, Lotufo RA (2016) Extinction profiles for the classification of remote sensing data. IEEE Trans Geosci Remote Sens 54(10):5631–5645

Giacinto G, Roli F (1997) Adaptive selection of image classifiers. In: International conference on image analysis and processing. Springer, Berlin, Heidelberg, pp. 38-45

Goldman S, Zhou Y (2000) Enhancing supervised learning with unlabeled data. In: ICML, pp 327-334

Goovaerts P (2011) A coherent geostatistical approach for combining choropleth map and field data in the spatial interpolation of soil properties. Eur J Soil Sci 62(3):371–380

Gu Y, Liu T, Jia X, Benediktsson JA, Chanussot J (2016) Nonlinear multiple kernel learning with multiple-structure-element extended morphological profiles for hyperspectral image classification. IEEE Trans Geosci Remote Sens 54(6):3235–3247

Guo H, Zhen L, Hao J, Wang C, Jie L, Liang D (2017) Big Earth Data: a new challenge and opportunity for Digital Earth's development. Int J Digit Earth 10(1):1–12

Hagenauer J, Omrani H, Helbich M (2019) Assessing the performance of 38 machine learning models: the case of land consumption rates in Bavaria, Germany. Int J Geogr Inf Sci 33(7):1399–1419

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer

Hermes L, Frieau D, Puzicha J et al (1999) Support vector machines for land usage classification in Landsat imagery. IEEE International Geoscience & Remote Sensing Symposium, pp 348–350

Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14(8):1771–1800

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507

Hsu CW, Lin CJ (2002) A comparison of methods for multiclass support vector machines. IEEE Trans Neural Netw 13(2):415–425

Hu F, Xia GS, Hu J, Zhang L (2015) Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens 7(11):14680–14707

Huang CQ, Davis LS, Townshend JRG (2002) An assessment of support vector machines for land cover classification. Int J Remote Sens 23(4):725–749

Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1–3):489–501

Huang B, Wu B, Barry M (2010a) Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. Int J Geogr Inf Sci 24(3):383–401

Huang SJ, Jin R, Zhou ZH (2010b) Active learning by querying informative and representative examples. In: Advances in neural information processing systems , pp 892-900

Inglada J (2007) Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. ISPRS J Photogramm Remote Sens 62(3):236–248

Kanevski M, Pozdnukhov A, Timonin V (2008) Machine learning algorithms for geospatial data. Applications and software tools. International Congress on Environmental Modelling and Software. 53. https://scholarsarchive.byu.edu/iemssconference/2008/all/53

Kanevski M, Vadim T, Alexi P (2009) Machine learning for spatial environmental data: theory, applications, and software. EPFL Press

Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK (2001) Improvements to Platt's SMO algorithm for SVM classifier design. Neural Comput 13(3):637–649

Kennedy J, Eberhart RC, Shi Y (2001) Swarm intelligence. Morgan Kaufmann Publishers, San Francisco

Ko AH, Sabourin R, Britto AS Jr (2008) From dynamic classifier selection to dynamic ensemble selection. Pattern Recogn 41(5):1718–1731

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097-1105

Kumar A, Kim J, Lyndon D, Fulham M, Feng D (2016) An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE J Biomedical Health Inform 21(1):31–40

Kuncheva LI (2002) Switching between selection and fusion in combining classifiers: an experiment. IEEE Trans Syst Man Cybern, Part B (Cybernetics) 32(2):146–156

Kuo BC, Ho HH, Li CH, Hung CC, Taur JS (2013) A kernel-based feature selection method for SVM with RBF kernel for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7(1): 317–326

Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE, Vol 2, pp 2169-2178

Leng J, Li T, Bai G, Dong Q, Dong H (2016) Cube-CNN-SVM: a novel hyperspectral image classification method. In: 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, pp 1027-1034

Li J, Heap AD, Potter A, Daniell JJ (2011a) Application of machine learning methods to spatial interpolation of environmental variables. Environ Model Softw 26:1647–1659

Li J, Bioucas-Dias JM, Plaza A (2011b) Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields. IEEE Trans Geosci Remote Sens 50(3): 809–823

Li E, Du P, Samat A, Meng Y, Che M (2016) Mid-level feature representation via sparse autoencoder for remotely sensed scene classification. IEEE J Sel Top Appl Earth Obs Remote Sens 10(3):1068–1081

Li E, Xia J, Du P, Lin C, Samat A (2017) Integrating multilayer features of convolutional neural networks for remote sensing scene classification. IEEE Trans Geosci Remote Sens 55(10):5653–5665

Lin Y, Kang M, Wu Y, Du Q, Liu T (2020) A deep learning architecture for semantic address matching. Int J Geogr Inf Sci 34(3):559–576

Liu W, Du P, Zhao Z, Zhang L (2016) An adaptive weighting algorithm for interpolating the soil potassium content. Sci Rep 6:23889

Liu W, Zhang HR, Yan DP, Wang SL (2017) Adaptive surface modeling of soil properties in complex landforms. ISPRS Int J Geo Inf 6(6): 178

Liu, W., Liu, Y., Yang, M., & Xie, M. (2018). Soil property surface modeling based on ensemble learning for complex landforms. In International Conference on Geo-informatics in Sustainable Ecosystem and Society (pp. 1-14). Springer, Singapore

Li Z, Chen J, Baltsavias E (2008) Advances in photogrammetry, remote sensing and spatial information sciences: 2008 ISPRS congress book. CRC Press

Luo T, Kramer K, Goldgof DB, Hall LO, Samson S, Remsen A, Hopkins T (2005) Active learning to recognize multiple types of plankton. J Mach Learn Res 6(Apr):589–613

Mamitsuka NAH (1998) Query learning strategies using boosting and bagging. In Machine learning: Proceedings of the Fifteenth International Conference (ICML'98) (Vol. 1). Morgan Kaufmann Pub

Marconcini M, Camps-Valls G, Bruzzone L (2009) A composite semisupervised SVM for classification of hyperspectral images. IEEE Geosci Remote Sens Lett 6(2):234–238

Melgani F, Bruzzone L (2004) Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans Geosci Remote Sens 42(8):1778–1790

Mountrakis G, Jungho I, Caesar O (2011) Support vector machines in remote sensing: a review. ISPRS Journal of Photogrammetry and Remote Sensing 66(3):247–259

Nawar S, Buddenbaum H, Hill J, Kozak J, Mouazen AM (2016) Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. Soil Tillage Res 155:510–522

Nogueira K, Penatti OA, dos Santos JA (2017) Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recogn 61:539–556

Ou D, Tan K, Du Q, Zhu J, Wang X, Chen Y (2019) A novel tri-training technique for the semi-supervised classification of hyperspectral images based on regularized local discriminant embedding feature extraction. Remote Sens 11(6):654

Pal M, Foody GM (2010) Feature selection for classification of hyperspectral data by SVM. IEEE Trans Geosci Remote Sens 48(5):2297–2307

Penatti OA, Nogueira K, Dos Santos JA (2015) Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 44-51

Perronnin F, Dance C (2007) Fisher kernels on visual vocabularies for image categorization. In 2007 IEEE conference on computer vision and pattern recognition, IEEE, pp 1-8

Platt JC, Cristianini N, Shawe-Taylor J (2000) Large margin DAGs for multiclass classification. In: Advances in neural information processing systems, pp 547-553

Rakotomamonjy A, Mary X, Canu S (2005) Non-parametric regression with wavelet kernels. Appl Stoch Model Bus Ind 21(2):153–163

Renda A, Barsacchi M, Bechini A, Marcelloni F (2019) Comparing ensemble strategies for deep learning: an application to facial expression recognition. Expert Syst Appl 136:1–11

Riitters KH, Wickham JD, Wade TG (2009) An indicator of forest dynamics using a shifting landscape mosaic. Ecol Indic 9(1):107–117

Rodriguez JJ, Kuncheva LI, Alonso CJ (2006) Rotation forest: a new classifier ensemble method. IEEE Trans Pattern Anal Mach Intell 28(10):1619–1630

Roli F, Fumera G (2001) Support vector machines for remote sensing image classification. Image & Signal Processing for Remote

Sensing VI. International Society for Optics and Photonics 4170: 160–166

Romero A, Gatta C, Camps-Valls G (2015) Unsupervised deep feature extraction for remote sensing image classification. IEEE Trans Geosci Remote Sens 54(3):1349–1362

Rossel RV, Behrens T (2010) Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158(1–2):46–54

Samat A, Du P, Liu S, Li J, Cheng L (2014) E2LMs: ensemble extreme learning machines for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7(4):1060–1069

Samat A, Gamba P, Du P, Luo J (2015) Active extreme learning machines for quad-polarimetric SAR imagery classification. Int J Appl Earth Obs Geoinf 35:305–319

Samat A, Gamba P, Liu S, Du P, Abuduwaili J (2016a) Jointly informative and manifold structure representative sampling based active learning for remote sensing image classification. IEEE Trans Geosci Remote Sens 54(11):6803–6817

Samat A, Li J, Liu S, Du P, Miao Z, Luo J (2016b) Improved hyperspectral image classification by active learning using pre-designed mixed pixels. Pattern Recogn 51:43–58

Samiappan S, Moorhead RJ (2015) Semi-supervised co-training and active learning framework for hyperspectral image classification. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, pp 401-404

Scudder H (1965) Probability of error of some adaptive pattern-recognition machines. IEEE Trans Inf Theory 11(3):363–371

Shi W, Zheng S, Tian Y (2009) Adaptive mapped least squares SVM–based smooth fitting method for DSM generation of LIDAR data. Int J Remote Sens 30(21):5669–5683

Shi W, Goodchild MF, Brian L, Yee L (2012) Advances in geo-spatial information science. CRC Press

Smits PC (2002) Multiple classifier systems for supervised remote sensing image classification based on dynamic classifier selection. IEEE Trans Geosci Remote Sens 40(4):801–813

Starck JL, Elad M, Donoho DL (2005) Image decomposition via the combination of sparse representations and a variational approach. IEEE Trans Image Process 14(10):1570–1582

Steiniger S, Taillandier P, Weibel R (2010) Utilising urban context recognition and machine learning to improve the generalisation of buildings. Int J Geogr Inf Sci 24:253–282

Fotheringham S, Crespo R, Yao J (2015) Geographical and temporal weighted regression (GTWR). Geogr Anal 47(4):431–452

Su H, Yong B, Du P, Liu H, Chen C, Liu K (2014) Dynamic classifier selection using spectral-spatial information for hyperspectral image classification. J Appl Remote Sens 8(1):085095

Tan K, Du P (2010) Classification of hyperspectral image based on morphological profiles and multi-kernel SVM. In: 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, IEEE, pp 1-4

Tan K, Du P (2011) Combined multi-kernel support vector machine and wavelet analysis for hyperspectral remote sensing image classification. Chin Opt Lett 9(1):011003–011006

Tan K, Li E, Du Q, Du P (2014) An efficient semi-supervised classification approach for hyperspectral imagery. ISPRS J Photogramm Remote Sens 97:36–45

Tan K, Hu J, Li J, Du P (2015) A novel semi-supervised hyperspectral image classification approach based on spatial neighborhood information and classifier combination. ISPRS J Photogramm Remote Sens 105:19–29

Tan K, Zhu J, Du Q, Wu L, Du P (2016) A novel tri-training technique for semi-supervised classification of hyperspectral images based on diversity measurement. Remote Sens 8(9):749

Tan K, Wang H, Zhang Q, Jia X (2018) An improved estimation model for soil heavy metal (loid) concentration retrieval in mining areas using reflectance spectroscopy. J Soils Sediments 18(5):2008–2022

Tarabalka Y, Benediktsson JA, Chanussot J, Tilton JC (2010a) Multiple spectral–spatial classification approach for hyperspectral data. IEEE Trans Geosci Remote Sens 48(11):4122–4132

Tarabalka Y, Chanussot J, Benediktsson JA (2010b) Segmentation and classification of hyperspectral images using watershed transformation. Pattern Recogn 43(7):2367–2379

Tuia D, Ratle F, Pacifici F, Kanevski MF, Emery WJ (2009) Active learning methods for remote sensing image classification. IEEE Trans Geosci Remote Sens 47(7):2218–2232

Tuia D, Volpi M, Copa L, Kanevski M, Munoz-Mari J (2011) A survey of active learning algorithms for supervised remote sensing image classification. IEEE Journal of Selected Topics in Signal Processing 5(3):606–617

Tumer K, Ghosh J (1996) Error correlation and error reduction in ensemble classifiers. Connect Sci 8(3–4):385–404

Vardhana M, Arunkumar N, Lasrado S, Abdulhay E, Ramirez-Gonzalez G (2018) Convolutional neural network for bio-medical image segmentation with hardware acceleration. Cogn Syst Res 50:10–14

Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning, ACM, pp 1096-1103

Vohland M, Besold J, Hill J, Fründ HC (2011) Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. Geoderma 166(1):198–205

Wan L, Hong Y, Huang Z, Peng X, Li R (2018) A hybrid ensemble learning method for tourist route recommendations based on geo-tagged social networks. Int J Geogr Inf Sci 32(11):2225–2246

Wang R, Chow CY, Kwong S (2015) Ambiguity-based multiclass active learning. IEEE Trans Fuzzy Syst 24(1):242–248

Wang Q, Gu Y, Tuia D (2016) Discriminative multiple kernel learning for hyperspectral image classification. IEEE Trans Geosci Remote Sens 54(7):3912–3927

Wang M, Zhang X, Niu X, Wang F, Zhang X (2019) Scene classification of high-resolution remotely sensed image based on ResNet. Journal of Geovisualization & Spatial Analysis 3(2):16

Woods K, Kegelmeyer WP, Bowyer K (1997) Combination of multiple classifiers using local accuracy estimates. IEEE Trans Pattern Anal Mach Intell 19(4):405–410

Wu F, Zhao Y (2006) Least squares support vector machine on gaussian wavelet kernel function set. In: International Symposium on Neural Networks, Springer, Berlin, Heidelberg, pp 936-941

Xia J, Du P, He X, Chanussot J (2014a) Hyperspectral remote sensing image classification based on rotation forest. IEEE Geosci Remote Sens Lett 11(1):239–243

Xia J, Chanussot J, Du P, He X (2014b) Spectral–spatial classification for hyperspectral data using rotation forests with local feature extraction and Markov random fields. IEEE Trans Geosci Remote Sens 53(5): 2532–2546

Xia J, Chanussot J, Du P, He X (2015a) Rotation-based support vector machine ensemble in classification of hyperspectral data with limited training samples. IEEE Trans Geosci Remote Sens 54(3):1519–1531

Xia J, Dalla Mura M, Chanussot J, Du P, He X (2015b) Random subspace ensembles for hyperspectral image classification with extended morphological attribute profiles. IEEE Trans Geosci Remote Sens 53(9): 4768–4786

Xia GS, Hu J, Hu F, Shi B, Bai X, Zhong Y, Zhang L, Lu X (2015) AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Trans Geosci Remote Sens 55(7):3965–3981

Xue Z, Du P, Su H (2014) Harmonic analysis for hyperspectral image classification integrated with PSO optimized SVM. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7(6):2131–2146

Xue Z, Li J, Cheng L, Du P (2015) Spectral–spatial classification of hyperspectral data via morphological component analysis-based image separation. IEEE Trans Geosci Remote Sens 53(1):70–84

Yang Y, Newsam S (2010) Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, pp 270-279

Yu J, Wang M, Tao D (2012) Semisupervised multiview distance metric learning for cartoon synthesis. IEEE Trans Image Process 21(11): 4636–4648

Yue TX, Wang SH (2010) Adjustment computation of HASM: a high-accuracy and high-speed method. Int J Geogr Inf Sci 24(11):1725–1743

Zhang L, Zhou W, Jiao L (2004) Wavelet support vector machine. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 34(1):34–39

Zhang F, Du B, Zhang L (2015) Scene classification via a gradient boosting random convolutional network framework. IEEE Trans Geosci Remote Sens 54(3):1793–1802

Zhao B, Zhong Y, Xia GS, Zhang L (2016) Dirichlet-derived multiple topic scene classification model fusing heterogeneous features for high spatial resolution remote sensing imagery. IEEE Trans Geosci Remote Sens 54(4):2108–2123

Zhong P, Wang R (2010) Learning conditional random fields for classification of hyperspectral images. IEEE Trans Image Process 19(7): 1890–1907

Zhou Y, Goldman S (2004) Democratic co-learning. In: 16th IEEE International Conference on Tools with Artificial Intelligence, IEEE, pp 594-602

Zhou ZH, Li M (2005) Tri-training: exploiting unlabeled data using three classifiers. IEEE Trans Knowl Data Eng 11:1529–1541

Zhu X, Goldberg AB (2009) Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning 3(1):1–130

Zou Q, Ni L, Zhang T, Wang Q (2015) Deep learning based feature selection for remote sensing scene classification. IEEE Geosci Remote Sens Lett 12(11):2321–2325