Journal of Geomatics Vol 11 No. 1 April 2017

Review Article

A survey of modern classification techniques in remote sensing for improved image classification

Mahmoud Salah
Department of Surveying Engineering, Shoubra Faculty of Engineering, Benha University, Egypt
Email: mahmoud.goma@feng.bu.edu.eg

(Received: Dec 13, 2016; in final form: Apr 13, 2017)

Abstract: Land Use and Land Cover (LULC) maps are the most important products of remote sensing which can be managed through a process called image classification. This paper reviews the major advanced classification approaches such as Artificial Neural Network (ANN), Classification Trees (CTs) and Support Vector machines (SVMs). This work compares performance of conventionally vis-à-vis recent classification techniques on satellite data. In addition, there are several issues requiring consideration in respect of the classification of remotely sensed data: 1) how to select the proper size of training samples? 2) how to set up the classifier parameters? and 3) how to combine classifiers in an efficient way? The objective of this paper is to answer these questions based on an intensive literature review. This review suggests that effective use of multiple features of remotely sensed data and the selection of a suitable classification method are pivotal for improving classification accuracy. More research, however, is needed to identify and reduce uncertainties in the image-processing to improve classification accuracy.

Keywords: Remote sensing, Image classification, Summary of reviews

1. Introduction

Till today, there is a need to produce regional Land Use and Land Cover (LULC) maps for a variety of applications such as landscape planning, change detection, disaster monitoring, resource management, site suitability analysis and ecological studies (Jensen, 2005). Remotely sensed images provide quantitative and qualitative information that reduces complexity and time of field work and can be used for producing LULC maps through a process called image classification (Chaichoke et al., 2011). Image classification is the process of extracting valuable information from massive satellite imagery by categorizing the image pixel values into meaningful categories or land cover classes. In the context of remote sensing, pixel is the ground area corresponding to one number of a digital image data set. The idea behind image classification is that different features on the earth's surface have a different spectral reflectance (Lillesand and Keifer, 2004).

With the advances of high resolution (HR) and very-high resolution (VHR) remotely sensed imagery such as IKONOS, QuickBird and World View, modern classification techniques are recently gaining the interest of the researchers. Comprehensive review of image classification techniques is required. Lu and Weng (2007) examined current practices, problems and prospects of image classification and summarized major advances in classification algorithms. Recently, Kumar and Singh (2013) reviewed digital image processing techniques for feature extraction from HR satellite imagery. Kamavisdar et al. (2013) have provided a brief theoretical knowledge about different image classification algorithms. Abburu and Golla (2015) summarized the various reviews on satellite image classification methods and techniques. Prasad et al. (2015) summarized the widely used advanced classification techniques that are used to improve classification accuracy. They considered various remote sensing features including spectral, spatial, multi temporal, multi sensor information, as well as ancillary data. Minu and Bindhu (2016) analyzed different supervised classification algorithms, post classification techniques and spectral contextual classification. The present review provides a comparative study on the efficiency, advantages and limitations of these techniques.

The motivation behind this review is to help the analyst, especially those who are new to the field of remote sensing, to select the most suitable classification approach in order to successfully classify a remotely sensed satellite imagery to produce a LULC map. In this review, recent advances in classification algorithms are considered such as Artificial Neural Network (ANN), Classification Trees (CTs) and Support Vector machines (SVMs). On the other hand, the most common problems associated with them have been discussed.

2. Remote-sensing classification process

According to Lu and Weng (2007), the major steps of image classification may include:

- Choice of a suitable classification system;
- Design image classes such as urban, agriculture, water areas, etc;
- Conduct field surveys and collect ground information;
- Image preprocessing for the enhancement of geometric and radiometric qualities of satellite imagery;
- Feature extraction and selection;
- Selection of training samples;
- Image classification;

- Post-processing: filtering, and classification decorating; and
- Accuracy assessment: compare classification results with ground truth data.

3. Selection of remotely sensed data

Remotely sensed data varies in spatial, spectral, temporal and radiometric resolutions. In order to get a better image classification, the most suitable sensor data should be selected. The characteristics of remotely sensed data are summarized by Lefsky and Cohen (2003). Many factors should be considered while selecting suitable sensor data such as scale, availability, characteristics, cost, time constraints and analyst's experience in using selected imagery. At a local level, HR data such as IKONOS and SPOT 5 data are the most useful data. At a regional scale, medium spatial resolution data such as Landsat TM/ETM+ and Terra ASTER are the most commonly used data. At a global scale, coarse spatial resolution data such as AVHRR, MODIS and SPOT Vegetation are needed (Lu and Weng, 2007). In general, spatial resolution is the most important factor that affects classification details and influences the selection of a classification algorithm as shown in table 1.

Table 1: Relation between spatial resolution and classification approach (Prasad et al., 2015).

High resolution	- Objects are made up of several pixels.
	- Object-based classification is superior to traditional pixel-based one.
Medium/low resolution	- Pixels and objects are similar in scale.
	- Both pixel-based and object-based image classifications perform well.

4. Data Preprocessing

It is necessary to check the quality of the remotely sensed data before stepping to classification stage. Image preprocessing includes restoration of bad lines; geometric rectification; radiometric calibration; and atmospheric and topographic corrections. If single data source is applied in classification, atmospheric corrections may not be required. If the study area includes rugged or mountainous regions, a topographic correction becomes necessary (Hale and Rock, 2003). A wide range of correction techniques are presented in Hadjimitsis et al. (2004). The detailed description of such corrections is beyond the scope of this review.

5. Feature extraction and selection

An effective use of features or attributes as input data for a classification procedure can improve the classification accuracy. A wide variety of variables are available which includes spectrum signature, vegetation indices, transformed images, textual information, height texture or surface roughness, multitemporal images, multisensor images, ancillary data (for non-spectral geographical information) and shape and size of objects. The selection of the most useful set of attributes for a classification process is necessary in order to reduce dimensionality of datasets without scarifying accuracy. On the other hand, it is necessary to compensate for some common problems associated with HR data such as shadows and the spectral variability within the same land-cover class (Lu and Weng, 2007). Many techniques can be applied for feature extraction which include principle component analysis (PCA), minimum noise fraction (MNF), transform discriminant analysis (TDA), decision boundary (DP), feature extraction (FE), non-parametric weighted feature extraction (NPWFE), wavelet transform (WT) and spectral mixture analysis (SMA). Table 2 summarizes the research efforts to improve the classification accuracy by applying such features in the classification process (Prasad et al., 2015):

Table 2: Using multiple features for improving classification accuracy

Method	Features	References
Ancillary data	DEM - land use - soil maps	(Maselliet al., 2000) (Baban and Yusof, 2001)
	Road density - road coverage - census data	(Zhang et al., 2002) (Epstein et al., 2002)
Stratification	Topography - census data - shape index of the Patches	(Bronge, 1999) (Helmer et al., 2000)
Post classification processing	Housing density - contextual correction	(Groom et al., 1996)
	Co-occurrence matrix - polygon and rectangular mode filters - expert system – knowledge based	(Zhang, 1999) (Stefanov et al., 2001) (Salah, 2014)
multisource data	Spectral – texture - ancillary	(Tso and Mather, 1999) (Trinder et al., 2010)

6. Selection of training samples

A better classification can be achieved only by considering a suitable classification algorithm with sufficient number of training samples. Training samples are often prepared by fieldwork or from other data sources such as aerial photographs and satellite imagery of fine spatial resolution based on single pixel, seed or polygon. In case of coarse resolution data, the selection of training samples is often tedious as it contains large regions of mixed pixels. Mixed pixels are due to existence of different classes in the same pixel. The purpose of generating training samples is to assemble a set of statistics that describe the spectral response patterns for each land cover class to be classified in the image (Lillesand and Kiefer, 2004). These training samples will be used later to train the algorithm. In case of parametric classifiers, for a fixed sample size, as the dimensionality of the data increases beyond a limit, the precision of the model parameter become lower (Hughes phenomenon). In this regard, it might be difficult to have a significant number of training pixels, and consequently parametric classifiers are not adequate to integrate ancillary data (Caetano, 2009). According to Kavzoglu and Mather (2003), the training sample sizes should range between [30 * Ni *(Ni + 1)] and [60 * Ni *(Ni + 1)] depending on the difficulty of the problem under consideration, where Ni is the number of input features or layers.

7. Classification approaches

There is a variety of classification techniques that have been developed and widely used to produce LULC maps. Satellite image classification methods can be broadly classified into three categories 1) unsupervised 2) supervised and 3) hybrid (Abburu and Golla, 2015). All three methods have their own advantages and disadvantages.

Unsupervised classification technique uses clustering mechanisms to group satellite image pixels into unlabeled classes/clusters. The analyst identifies the number of classes/clusters to generate and which bands to use. Based on this information, the image classification algorithm generates classes/clusters. In order to produce well classified satellite image, the analyst manually identifies each cluster labels a land cover class. It is often the case that multiple clusters represent a single land cover class. The analyst merges clusters into a single land cover class. The unsupervised classification technique is commonly used when no training sample sites exist. There are two most frequent clustering methods used for unsupervised classification, namely, K-means and Iterative Self-Organizing Data Analysis Technique (ISODATA). These two methods rely purely on pixel-based statistics and incorporate no prior knowledge of the characteristics of the themes under investigation.

On the other hand, supervised classification is a method in which the analyst defines small representative samples for each land cover class called training sample. In supervised classification, the analyst must be familiar with the area covered by the satellite image and the spectral properties of the land cover classes. Accuracy of the classification results highly depends on the samples taken for training. The image classification algorithm uses the training samples to identify the land cover classes in the entire image. The common supervised classification algorithms are minimum distance (MD), Mahalanobis distance (MhD), parallelepiped (PP), maximum likelihood classifier (MXL), K-nearest neighbor (KNN), SVMs, and spectral angle mapper (SAM) (Jawak et al., 2015). Figure 1 shows the major steps in the two major types of image classification (Al-doski et al., 2013):

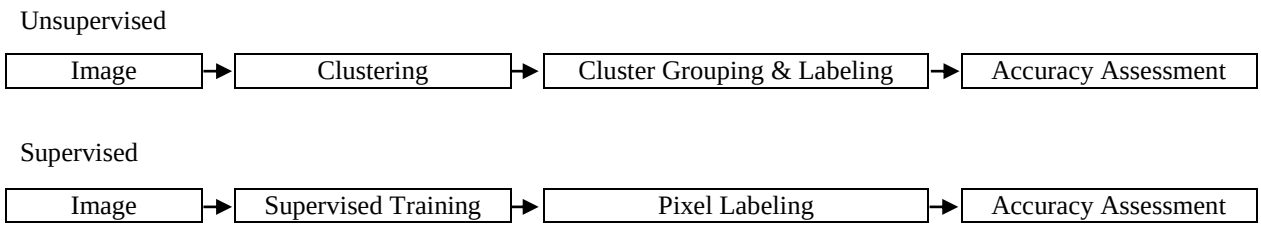

Figure 1: The major steps of supervised and unsupervised image classification

Supervised and unsupervised classifications can be used as alternative approaches, but are often combined to form a hybrid system using more than one methods. On the other hand, when using new generation of images, characterized by a higher spatial and spectral resolution, it is still difficult to obtain satisfactory results by using supervised and unsupervised techniques alone. More specifically, a wide variety of classification categories is available. For the sake of convenience, this review categorized classification approaches as shown in table 3.

Table 3: A taxonomy of image classification methods (Kamavisdar et al., 2013)

Criteria	Categories	Characteristics	Example
Training Sample	Supervised	- Use training sets to classify pixels of unknown identity.	- MD - PP - MXL
	Unsupervised	- Divides pixels into number of classes based on natural groupings. - No prior knowledge is required.	- K-means - ISODATA
Assumptions on Data distribution	Parametric	- Based on assumption of Gaussian distribution. - Mean vector and covariance matrix are generated from training samples.	- MXL
	Non-Parametric	- No prior assumptions about data distribution.	- ANN - SVMs - CTs - Expert system - Knowledge based
Number of Outputs	Hard (crisp)	- Each pixel shows membership to single class.	- MXL - MD - ANN - CTs - SVMs
	Soft (fuzzy)	- Each pixel exhibits partial class membership. - Produces more accurate result. - Ability to deal with mixed pixels.	- MXL - ANN - Fuzzy C-means (FCM)
Pixel Information	Per-pixel classifier (PP)	- Pixel by pixel classification. - Generates signatures by using the spectra of all training pixels. - Low accuracy because of the impact of mixed pixel problem. - Salt and pepper phenomenon.	- MXL - ANN - SVMs - MD
	Sub-pixel classifiers	- Provides membership of each pixel to each class. - Has the capability to handle the mixed pixel problem. - Suitable for medium and coarse spatial resolutions. - Difficult to access accuracy.	- SMA - Fuzzy classifiers
	Per-field	- Averages out the noise by using land parcels as individual units. - Integrates vector and raster data. - Difficult to handle the dichotomy between vector and raster data. - Suitable for fine spatial resolutions	- GIS-based approaches
	Object-oriented (OO)	- Pixels are grouped into objects of different shape and scale (segmentation) and then classification is performed on the basis of objects. - Additional information such as object texture, shape and relations to adjacent regions can be used. - Perfect especially for HR imagery. - Over- and under-segmentation may reduce the classification accuracy.	e-Cognition software
Spatial Information	Spectral	- Based on pure spectral information	- MXL - MD - ANN
	Contextual	- Spatial measurements related to the neighborhoods	- Markov random field
Multiple classifiers	Hybrid Systems	- combine the advantages of multiple classifiers	- Voting rules - Bayesian formalism - Evidential reasoning - Multiple ANN

8. Selection of suitable classification method

8.1 Classic classifiers

In addition to the aforementioned categories, this work has further categorized classifiers as classic and advanced classifiers. Most classic classifiers are based on assumptions of data distribution. The performance of such classifiers depends largely on the accuracy of the estimated model parameters. Classic classifiers suffer from the curse dimensionality of new satellite imagery (Hughes phenomenon). As a result, it might be difficult to select a significant number of training samples. Another drawback of the classic classifiers is the difficulty of combining spectral data with ancillary data (Wilkinson, 2005). Classic classifiers include ISODATA, K-Means, KNN, MD, MhD, PP, MXL and SAM. They are not discussed, since the readers can find them in many textbooks (Lillesand and Keifer, 2004). MXL, however, is the most widely used statistical supervised classifiers. This classifier is based on the Bayesian theory of probability and uses an array of patterns and a covariance matrix from a Gaussian distribution sample set. MXL allocates pixels to appropriate classes based on probability values of the pixels and has been adapted as an indicator of sub-pixel proportions. While using the MXL algorithm, several issues must be taken into consideration: 1) sufficient ground truth data should be sampled to allow accurate estimation of the mean vector and the variance-covariance matrix; 2) the inverse matrix of the variance-covariance matrix becomes unstable in the case of high correlation between two image bands; and 3) when the population is not normally distribution, the MXL algorithm cannot be applied (Kussul et al., 2006). Table 4 summarises the advantages and disadvantages of classic classifiers (Richards, 2013).

Table 4: Advantages and disadvantages of classic classifiers

Classifier	Advantages	Disadvantages
ISODATA	fast and simple to process	- Needs several parameters
K-Means	- Fast and simple to process	- Could be influenced by: the number and position of the initial cluster centers specified by the analyst, the geometric properties of the data, and clustering parameters
KNN	- Simple to process	- Computationally expensive when the training dataset is large
MD	- Fast and simple to process	- Considers only mean value
MhD	- Fast and simple to process	- Considers only mean value
PP	- Fast and simple to process	- Overlap may reduce the accuracy of the results
MXL	- Sub-pixel classifier	- Time consuming - insufficient ground truth data and/or correlated bands can affect the results - Cannot be applied when the dataset is not normally distribution

8.2 Advanced classification algorithms

The improvement in the spatial resolution and quality of remotely-sensed data does not guarantee more accurate feature extraction. The image classification techniques used are a very important factor for better accuracy (Robert et al., 2010). The advanced classification algorithms such as ANN, SVMs and CTs algorithms are highly applied for image classification and have commonly outperformed conventional classifiers in their performance. They are especially suitable for the incorporation of non-spectral data into the classification process. A brief description of each classifier is provided below. Readers who wish to have a detailed description of a specific classifier can refer to cited references.

8.2.1 Artificial Neural Networks (ANN) ANN is a form of artificial intelligence that simulates some functions of the human brain to associate the correct meaningful labels to image pixels. ANN-based classification uses nonparametric approach and hence it is easy to incorporate supplementary data in the classification process in order to improve classification accuracy (Abburu and Golla, 2015). An ANN consists of a series of layers, each containing a set of processing units called neurons. All neurons on a given layer are linked by weighted connections to all neurons on the previous and subsequent layers. During the training phase, the ANN learns about the regularities present in the training data and then constructs rules that can be extended to the unknown data (Foody, 1999). ANN algorithms are extremely efficient when the classification process is not a simple one. A well trained network is capable of classifying highly complex data.

There are several ANN algorithms that can be used to classify remotely sensed images which include:

8.2.1.1 Multi-layer perceptron (MLP): MLP is the most widely used type of ANN. It is a feed-forward ANN model that maps input data sets onto a set of appropriate outputs (Rosenblatt, 1962). MLP has three primary layers: input layer; output layer; and one or more hidden layers with each layer connected to the next one as shown in figure 2. Each layer is composed of a user-defined number of neurons. Input layer neurons represent the input variables while output layer neurons represent the classes specified by input training samples. In this regard, there is one input neuron for each input variable and one output layer neuron for each class. MLP utilizes a supervised learning technique called back-propagation for training the network. Mathematically this can be expressed as:

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) = \varphi(w^T x + b) \quad (1)$$

where w refers to the vector of weights, x is the vector of inputs, b is the bias and φ is the activation function. The activation function is normally selected to be the sigmoid 1/ (1+e-x). This function has proved to model nonlinear mappings well (Cybenko, 1989). MLP interprets the weights and activation functions of the neurons. Input and hidden layer neurons are randomly weighted and each pixel in the training data is assigned probability to an output neuron based on maximum activation. Each solution is compared with the previous one, and the solution that results in the lowest error is retained. The process continues until acceptable testing error for the partition of input variables into the specified output classes is reached. The trained network is then used to classify the remaining dataset based on the level of output neuron activation produced by a given pixel (Foody, 1995). The main difficulty with MLP is that it requires a complete retraining of the whole network. This may modify or even erase previous learning, and lead to longer training time even for small size dataset (Liu et al., 2004). In order to improve the MLP performance without costs large computation time, Kavzoglu and Mather (2003) have suggested a set of parameter values for MLP classifiers as shown in table 5 where N is the number of classes.

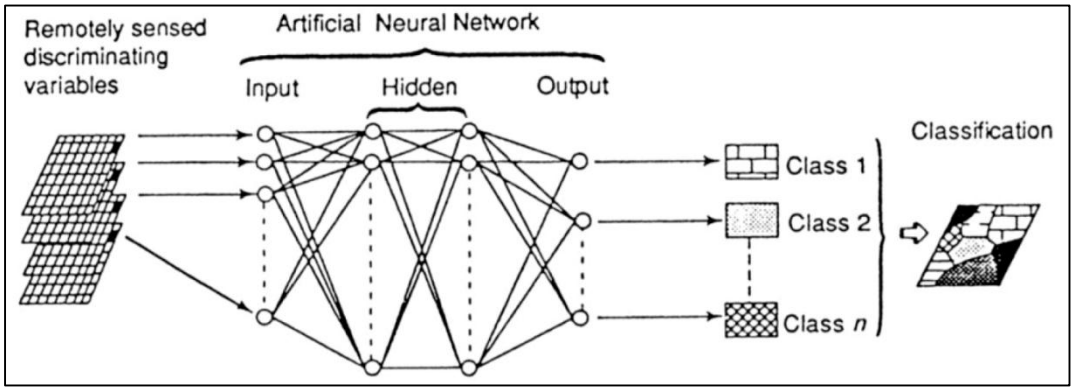

Figure 2: A typical MLP with back-propagation (Foody, 1999)

Table 5: The basic architecture to start MLP classifier

Number of hidden layers/nodes	Between $2N_i$ to $3N_i$
Learning rate	0.01- 0.001
Momentum factor	0.5
Sigmoid constant	1
RMSE	0.0001
Number of iterations	10000
Accuracy rate percent	100%

8.2.1.2 Fuzzy ArtMap classification: Fuzzy ArtMap performs classification based on Adaptive Resonance Theory (ART) (Carpenter et al., 1991). Fuzzy ArtMap is a clustering approach that operates on vectors with fuzzy inputs (real numbers between 0 and 1) and incorporates an incremental learning method to learn continuously without forgetting previous learned states (Oliveira et al., 2007). It adopts only the weights of the neurons encoding the class that best matches the input pattern. In this regard, it can solve large scale problems through a few training epochs. On the other hand, it is sensitive to noise and outliers that may lead to increased misclassified pixels. Fuzzy ArtMap consists of four layers of neurons: input (F1), category (F2), map field and output. Five parameters should be specified for the Fuzzy ArtMap as shown in table 6 (Li et al., 2012):

Table 6: The proposed parameters to start Fuzzy ArtMap classifier

Choice parameter α	A small positive constant
Learning rate parameters β1 in ARTa	$0 \leq \beta 1 \leq 1$
Learning rate parameters β2 in ARTb	$0 \leq \beta 2 \leq 1$
Vigilance parameters ρ1 in ARTa	Normally set very close to 1
Vigilance parameters ρ2 in ARTb	Normally set very close to 1

The ρ1 and ρ2 are the most important parameters and control the process during learning and operational phases of the network. Map field and category layer weights are learned adaptively during the process. Each input layer (F1) observation (pixel) is assigned to a category layer (F2) neuron based on its spectral data characteristics. If no F2 neuron meets the similarity threshold of a given F1 observation, a new F2 Neuron is created in order to partition subsets of a degree of homogeneity defined by the user through a vigilance parameter (Tso and Mather, 2009).

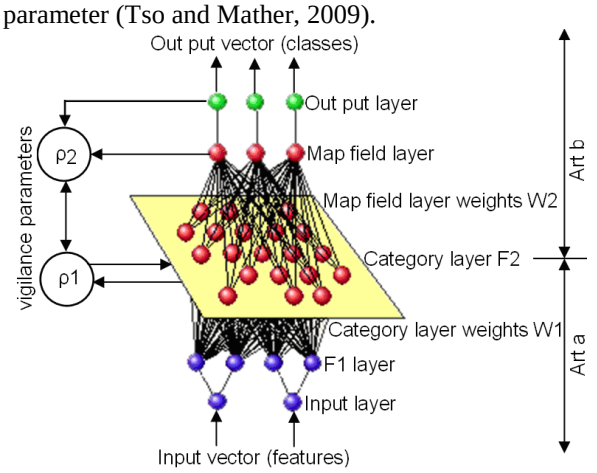

Figure 3: Fuzzy ArtMap architecture (Eastman, 2006)

8.2.1.3 Self-Organized feature Map (SOM) : SOM is a neural network algorithm composed of a single layer of neurons as shown in figure 4 (Kohonen, 1990). The input layer represents the input feature vector and thus has neurons for each measurement dimension. For the output layer of an SOM, a 15 x 15 array of neurons has been recommended by Hugo et al. (2007). Small networks of neurons may result in some unrepresented classes in the final labeled network. On the other hand, large arrays of neurons lead to improved overall classification accuracy. Synaptic weights that connect output layer neuron to all neurons in the input layer are randomly initialized and subsequently organized by systematic sampling of the input data. The organization process progressively adjusts the weights based on data characteristics such that neurons with similar weights spatially cluster in the neuron layer.

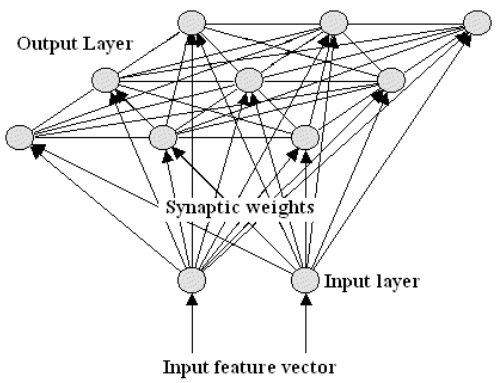

Figure 4: Example of a SOM with a 2 neurons input layer and 3x3 neurons output layer

During the training phase, each neuron with a positive activity within the neighborhood of the winning neuron participates in the learning process. A winning processing element is determined for each input vector based on the similarity between the input vector and the weight vector (Jen-Hon and Din-Chang, 2000). Let X= (x1, x2, x3…, xn) be a vector of reflectance for a single pixel input to the SOM. First, synaptic weights between the output and input neurons are randomly assigned (0-1). The distances between a weight vector and an input feature vector are then calculated, and the neuron in the output layer with the minimum distance to the input feature vector, winner neuron, is then determined. The weight of the winner and its neighbors within a radius γ are then altered (while those outside were left unaltered) according to a learning rate αt.

SOM supervised classification has two training phases: 1) unsupervised classification phase in which competitive learning and lateral interaction lead to a regional organization of neuron weights (topology); and 2) refinement of the decision boundaries between classes based on the training samples using a learning vector quantization (LVQ) algorithm (Nasrabadi and Feng, 1988). Each pixel is then assigned a class of the most similar neuron or neurons in weight (minimum Euclidian distance) to the pixel vector of reflectance. Unlike MLP or Fuzzy ArtMap, SOM acknowledges relationships between classes (i.e., feature map neurons), which allows for the discrimination of multimodal classes. On the other hand, the system normally yields many unclassified pixels (Qiu and Jensen, 2004). In order to improve the classification accuracy without costs large computation time, Vesanto et al. (2000) has suggested a set of parameter values for an SOM classifier as shown in table 7.

Table 7: The proposed parameters to start SOM classifier.

Course tuning parameters				**Fine tuning parameters**		
Output neurons	Initial γ	Min. α	Max. α	Min. δ^t	Max. δ^t	Fine tuning epoch
225 (15*15)	25	0.5	1	0.0001	0.0005	50

8.2.1.4 Radial Basis Function Network (RBFN) : RBFN is a non-linear neural network classifier that consists of an n-dimensional input vector, a layer of RBF neurons and an output layer with one node per category or class of data. An RBFN performs classification by measuring the similarity of input to training data. Each RBFN neuron stores a prototype, one example from the training set. A fairly straight forward approach for making an intelligent selection of prototypes is to perform k-Means clustering on the training set and to use the cluster centers as the prototypes. Each neuron computes the Euclidean distance between the input and its prototype and outputs a value, called activation value, between 0 and 1 which is a measure of similarity. If the input is equal to the prototype, then the output of that RBF neuron will be 1.

As the distance between the input and prototype grows, the response falls off exponentially towards 0. Each output node computes a sort of score for the associated category. The score is computed by taking a weighted sum of the activation values from every RBF neuron, and multiplies the neuron's activation by this weight before adding it to the total response. Typically, a classification decision is made by assigning the input to the category with the highest score.

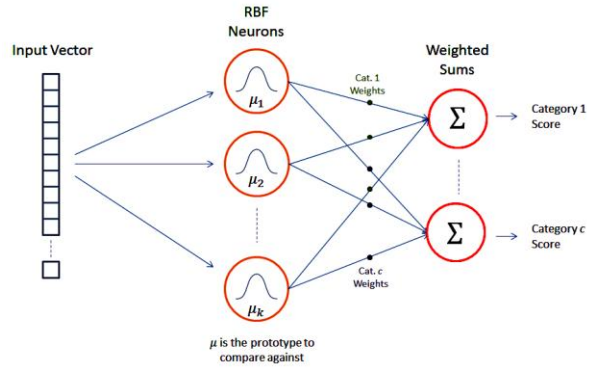

Figure 5: RBF Network Architecture

There is a variety of similarity functions, but the commonly used one is based on the Gaussian. Equation 2 represents a Gaussian with a one-dimensional input, where x is the input, μ is the mean, and σ is the standard deviation. The RBF neuron activation function is slightly different as shown in equation 3. The training process for an RBFN consists of selecting three sets of parameters: the prototypes (μ); β coefficient for each of the RBF neurons; and the matrix of output weights between the RBF neurons and the output nodes. In order to improve the classification accuracy, Hwang and Bang (1997) suggested setting the parameters μ and β to 1.05 and 5 respectively.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

$$\varphi(x) = e^{-\beta\|x-\mu\|^2} \quad (3)$$

8.2.2 Classification trees (CTs) : The theory of CT was introduced by Breiman et al. (1984). CT is a non-parametric, iterative and progressive method of pattern recognition based on hierarchical rule approach. A CT consists of the following elements: the root node (the starting node); the non-terminal nodes (between the root node and all other internodes); and the terminal node (that represents the group of pixels that are assigned to the same class as shown in figure 6. It predicts class membership by recursively partitioning a dataset into homogeneous subsets using a variety of binary splitting rules (Tso and Mather, 2009). These rules are derived from training data using statistical methods and based on the 'impurity'. If all pixels contained by a given node belong to the same category, the node is pure and the impurity is 0. If the logical condition at a given node is fulfilled, the left branch is chosen; otherwise the branch to the right is followed. The process continues until the node becomes pure and is assigned as a terminal node.

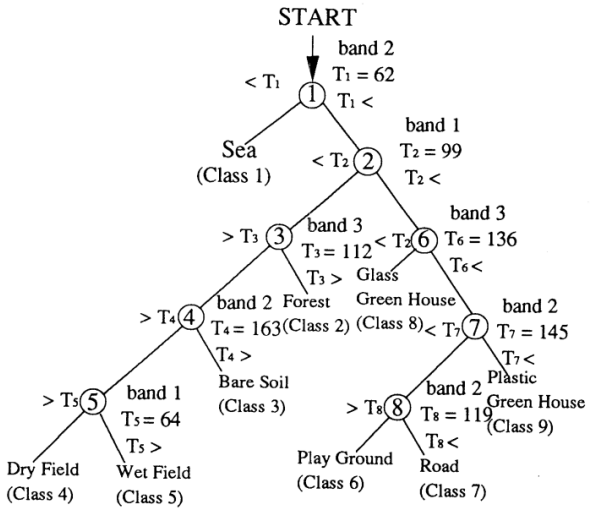

Figure 6: Classification tree. The numbers indicate the variables and their values that are used as thresholds for each node condition.

The most widely used splitting rules are: 1) the Entropy; the Gain Ratio or Information Gain (IG); and the Gini models. Entropy measures the homogeneity and aims to decrease the entropy until a pure terminal node, has zero entropy, is reached (Shannon, 1949). IG is a measure of reduction in Entropy that would result from splitting node N using rule T (Quinlan, 1987). By calculating IG(T) for each variable, the variable that achieves the highest IG will be chosen to split the data at that node. One drawback of this approach is that the variables with relatively high variances are generally selected. This would lead to a bias towards a large number of splits. In order to overcome this problem, the IG(T) can be adjusted by the entropy of the partitioning. The Gini index measures the impurity of the node and separates the largest homogeneous group from the remaining training data (Breiman et al., 1984). The Gini index of all parts is summed for each split rule. The split rule with the maximum reduction in impurity, minimum Gini index, is selected.

When the CT characterizes too much details or noise in the training data, an over-fitting process may occurs and reduces classification accuracy. Pruning normally results in small and more effective trees by up to 25% and avoids such fitting process. Among the proposed pruning methods, the 10-fold cross validation process has proved to be a robust method and does not require any independent dataset to assess the performance of the splitting model. The pruned tree is normally resulted in the best classification accuracy. More details about the cross-validation process are given by Sherrod (2008).

8.2.3 Support Vector Machines (SVMs) : SVMs are one of the more recent developments in the field of machine learning and based on the principles of

statistical learning theory (Vapnik, 1979). Mountrakis et al. (2011) summarized results from over 100 articles using the SVMs algorithm. In conclusion, SVMs have proved to be superior to most other image classification algorithms in terms of classification accuracy. SVMs as binary classifier delineate two classes by fitting an optimal separating hyperplane to the training data in the multidimensional feature space to maximize the margin between them. In figure 7, m is the distance between H1 and H2, and H is the optimum separation plane. For a binary classification problem in n-dimensional feature space, xi is a training set of l samples, i=1,2,…,l, with their corresponding class labels yi {1, -1}. The optimum separation plane is defined by equation 4, where x is a point on the hyperplane, w is an n-dimensional vector perpendicular to the hyperplane, and b is the distance of the closest point on the hyperplane to the origin. Equation 5 and equation 6 can be combined into equation 7. SVMs attempt to find a hyperplane, equation 4, with minimum ||w||2 that is subject to constraint 7.

$$w.x + b = 0 \quad (4)$$

$$w.x_i + b \le -1, \quad for\ class\ 0 \quad (5)$$

$$w.x_i + b \ge -1, \quad for\ class\ 1 \quad (6)$$

$$y_i\left[(w.x_i) + b\right] - 1 \ge 0 \quad \forall i \quad (7)$$

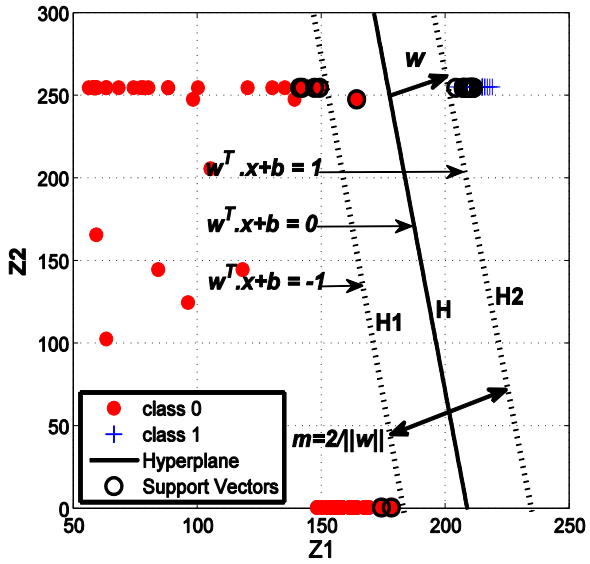

Figure 7: Optimum separation plane in the (Z1, Z2) space

Four kernel functions (functions used to project the data from input space into feature space) are available for SVMs: Gaussian Radius Basis Function (RBF); Linear; Polynomial; and Sigmoid (Quadratic). In remote sensing applications the Gaussian radial basis function (RBF) kernel has proved to be effective with reasonable processing times (Van der Linden et al., 2009). Two parameters should be specified while using RBF kernels: C, the penalty parameter that controls the trade-off between the maximization of the margin between the training data vectors and the decision boundary plus the penalization of training errors, and γ, the width of the kernel function. The problem is that C and γ depend on the data range and distribution and they differ from one classification problem to another. The most common used way to optimize the C and γ parameters is a grid-search using a 10-fold cross-validation error as a measure of quality. This method can prevent the overfitting problem and results in better accuracy (Hsu et al., 2009).

In order to solve for the binary classification problem that exists with SVMs and to handle the multi-class problems in remote sensing applications, two approaches are commonly used: the One-Against-All (1AA); One-Against-One (1A1). Anthony et al. (2007) have reported that the resulting classification accuracy from 1AA is not significantly different from 1A1 approach. However, the 1A1 technique results in a larger number of binary SVMs and then in subsequently intensive computations than the 1AA technique. The original output of a SVM represents the distances of each pixel to the optimal separating hyperplane, referred to as rule images. All positive (+1) and negative (-1) votes for a specific class are summed and the final class membership of a certain pixel is derived by a simple majority voting.

8.2.4 Fuzzy Classifiers : Fuzzy classifiers express the fuzzy set membership of each pixel in each class. The fuzzy set membership is calculated based on standardized Euclidean distance from the mean of the signature, using a specific algorithm. The underlying logic is that the mean of a signature represents the ideal point for the class, where fuzzy set membership is 1. When distance increases, fuzzy set membership decreases, until it reaches the user-defined distance where fuzzy set membership decreases to 0. The FCM clustering algorithm (Bezdec, 1981) is the most representative fuzzy classification algorithms since it is suitable for tasks dealing with overlapping clustering. The classification is performed with an iterative optimization of minimizing a fuzzy objective function (Jm) defined as equation 8.

$$J_m = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m d^2(x_k, V_i) \quad (8)$$

where
c = number of clusters
n = number of pixels
μik= membership value of ith cluster of kth pixel
m = fuzziness for each fuzzy membership.
xk= vector of kth pixel
Vi= center vector of ith cluster
d2(xk,Vi) = Euclidean distance between xkand Vi

The membership (μik) is estimated by the distance between kth pixel and center of ith cluster, and is constrained as follows:

$$\begin{cases} 0 \le \mu_{ik} \le 1 & \text{for all } i,k \\ \sum_{i=1}^{c}\mu_{ik} = 1 & \text{for all } k \\ 0 < \sum_{k=1}^{n}\mu_{ik} < n & \text{for all } i \end{cases} \quad (9)$$

The center of cluster (Vi) and the membership value (μik) could be calculated by equations 10 and 11, respectively.

$$V_i = \frac{\sum_{k=1}^{n}(\mu_{ik})^m x_k}{\sum_{k=1}^{n}(\mu_{ik})^m}, 1 \le i \le c \quad (10)$$

$$\mu_{ik} = \left[\sum_{j=1}^{c}\left(\frac{d(x_k,V_i)}{d(x_k,V_j)}\right)^{\frac{2}{m-1}}\right]^{-1}, 1 \le i \le c, 1 \le k \le n \quad (11)$$

Therefore, Jm can be minimized by iteration through equations 10 and 11. The first step of the iteration is to initialize a fixed c, a fuzziness parameter (m), a threshold ε of convergence, and an initial center for each cluster, then computing μik and Vi using equations 10 and 11 respectively. The iteration is terminated when the change in Vi between two iterations is smaller than ε. Finally, each pixel is classified into a combination of memberships of clusters.

Table 8: Comparison of modern classification techniques (Kamavisdar et al., 2013)

Method	Advantages	Disadvantages
ANN	- Non-parametric classifiers. - High computation rate of very large datasets - Efficiently handles noisy inputs	- It is difficult to understand how the result was achieved. - The training process is slow. - Problem of over fitting. - Difficult to select the type network architecture. - Dependent on user-defined parameters.
CTs	- Non-parametric classifiers - Does not require an extensive design and training. - Easy to understand the classification process. - Accurate and computational efficiency is good. - Easy to integrate multi-source data.	- Calculation becomes complex when various outcomes are correlated.
SVMs	- Non-parametric classifiers - Provides a good generalization. - The problem of over fitting is controlled. - Computational efficiency is good. - perform well with minimum training set size and high-dimensional data - Often outperform other classifiers.	- Training is time consuming. - Difficult to understand its structure. - Dependent on user-defined parameters. - Determination of optimal parameters is not easy.
Fuzzy Classifiers	- Efficiently handle overlapping data. - Minimize computation time and reduces memory requirements.	- Without priori knowledge output is not good

For a specific dataset, it is often difficult to identify the classifier with the best performance due to the lack of a guideline for selection on hand. Moreover, the analyst has to make many decisions and choices through image classification process. Many researchers have compared unsupervised, supervised and hybrid classification techniques. Table 9 provides summary of different researchers' conclusion and the situation in which each classifier is most useful. The researchers' opinion about the best classification method is not consistent. Many more suggestions on the selection of classifiers can be found in Foody et al. (2007)

Table 9: performance evaluation of various classification methods against different datasets

Researcher	Classifier	Datasets	Best Performance
Pal and Mather (2005)	- SVMs - MXL - ANN	- Landsat 7 ETM+ - Hyperspectral data	SVMs
Oliveira et al. (2007)	- MXL - CTs - MLP - SOM - ArtMap	- Landsat (ETM+)	ArtMap
Lippitt et al. (2008)	- SOM - MLP - ArtMap - CTs	- Landsat7 (ETM+)	CTs
Li et al. (2012)	- MXL - CTA - ArtMap - KNN	- Landsat 5 TM - ALOS PALSAR (L-band HH and HV)	ArtMap
Maryam et al. (2014)	- SVMs - MXL - MD - PP	- Landsat7 ETM+	SVMs
Shaker et al. -2012	- Contextual - MXL - MD	- SPOT	MXL
Mannan et al. (1998)	- ArtMap - MXL - MLP	- IRS-1B	ArtMap
Gil et al. -2011	- SVMs - ANN - MXL	- IKONOS	SVMs
Du et al. (2012)	- MLP - ANN - CTs - MXL - SVMs - ArtMap	- QuickBird - Spectral/textural features	SVMs
Doma et al. (2015)	- PP - MD - MXL - ANN - SVMs	- Quick bird	SVMs
Hamedianfar et al. (2014)	- OO/fuzzy - SVM	- World View-2 (WV-2)	OO/fuzzy
Camps-Valls et al. (2003)	- SVMs - ANN	- hyperspectral data (128 bands)	SVMs
Trinder et al. (2010)	- SVMs - SOM - CTs	- Aerial Images - LiDAR data	SVMs

9. Hybrid Classifiers

Different classifiers resulted in different classes for the same test area. No single classifier can performs the best for all classes. Many of the classification algorithms are complementary. Analyses of the results reported in Kanellopoulos et al. (1997) have confirmed the complementary information of neural and statistical algorithms. These classifiers result in uncorrelated classification errors and hence higher classification accuracies can be obtained by combining them. In the hybrid classifiers-based approach, the classifiers should use independent feature set and/or be trained on separate sets of training data. Two strategies exist for combining classifiers: 1) Classifier Ensembles (CE); and 2) Multiple Classifier Systems (MCS) as shown in figure 8.

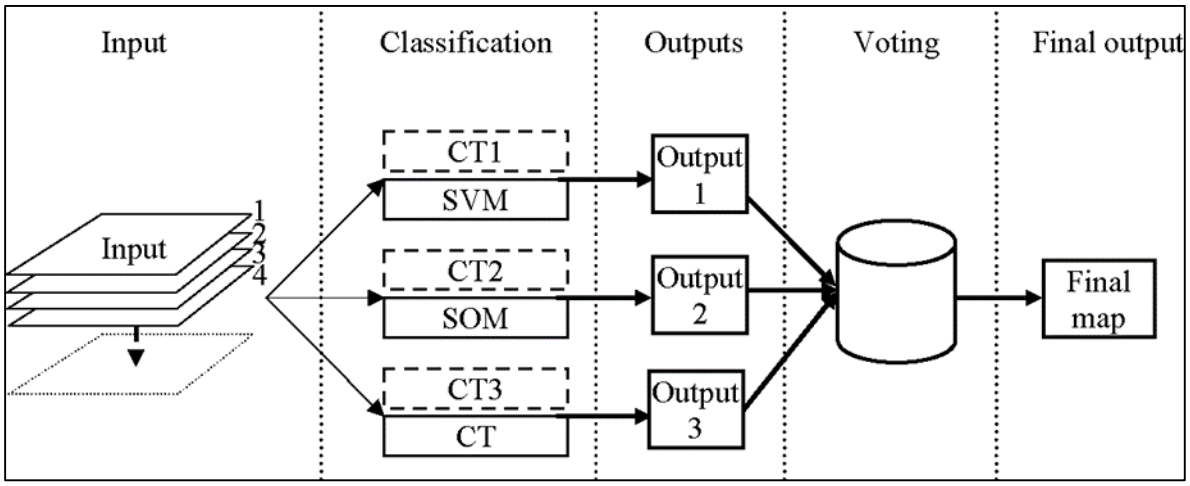

Figure 8: Classifier Ensemble (dashed) versus Multiple Classifier Systems (solid) (Waske, 2007, modified)

If the classification results are similar, the combination process would not improve the classification accuracy. Therefore, diversity is an important requirement for the success of hybrid systems (Chandra and Yao, 2006). Diversity measures are rarely used and compared for remote sensing image classification which includes: Kappa statistics (Congalton and Mead, 1983); double fault (Giacinto and Roli, 2001); agreement measure (Michail et al., 2002); similarity, non-inferiority, difference (Foody, 2009); weighted count of errors and correct results (WCEC) (Aksela and Laaksonen, 2006), entropy (Kuncheva and Whitaker, 2003); and Disagreement-accuracy measure (Du et al., 2012). The results obtained by Du et al. (2012) indicate that the combination selected by Disagreement accuracy measure outperform the ones selected by other diversity measures in terms of overall accuracy.

9.1. Classifier Ensembles

Classifier ensembles are based on the combination of a variety of the same algorithm. By training the so-called base classifier on modified training data, a set of independent classifiers can be obtained. Normally, a voting scheme is applied to combine the results. The widely applied strategies for generating classifier ensembles include: 1) resampling of the training data, such as bagging or boosting; and 2) resampling of the input features, such as random feature selection (Waske, 2007).

9.1.1. Bagging (bootstrap aggregating) or Boosting :

Bagging (Breiman, 1996) and boosting (Freund and Schapire, 1997) have been proposed to combine classifiers. The performance of such algorithms is limited by the high level of ambiguities among classes which may result in poor classification accuracy (Yu-Chang and Kun-Shan, 2009). In bagging, n samples are selected randomly from a training set with k bags, created iteratively, and each bag is classified by vote to predict its class. Some training instances will occur multiple times in a bag, while others may not appear at all. After that, each bag is used to train a classifier. These classifiers are then combined. Such a method is not very sensitive to noise in the data. On the other hand, boosting is based on multiple learning iterations. At each iteration, instances that are incorrectly classified are given a higher weight in the next iteration. By doing so, the classifier is forced to concentrate on instances that were not correctly classified in earlier iterations. At the end, all of the trained classifiers are combined.

Bagging has proved to reduce the variance of the classification, while Boosting reduces both the variance and the bias of the classification. In this regard, Boosting can produce more accurate classification results than Bagging (Du et al., 2012). However, Boosting is computationally more demanding than other simpler algorithms, while the lack of robustness to noise is another shortcoming (Benediktsson et al., 2007). In addition, there is a great variety of approaches drawn upon the basic idea of Bagging and Boosting such as: Wagging (Bauer and Kohavi, 1999), Random Forest (Breiman, 2001), Random Subspace (Ho, 1998), Logistic Boosting (Collins et al., 2002), MultiBoost (Webb, 2000), Rotation Forest (Rodriguez and Kuncheva, 2009), and Rotboost (Zhang and Zhang, 2008).

9.1.2. Random Feature Selection (RFS) : Another approach for generating independent classifiers is the modification of the input feature space, by a random selection of features. This approach has proved to be superior to bagging and boosting, since the method normally selects a subset of the available input features without replacement. The number of selected features within the subset is user-defined, and is usually set to the square root of the number of input features. The computational cost is lighter than bagging and boosting because it is only based on subsets of input data. Because of that this method can handle high-dimensional data sets. On the other hand, the correlation between the classifiers is decreased, and the obtained classification accuracy is improved (Gislason et al. 2006).

9.2. Multiple Classifier Systems (MCS)

In contrast to the classifier ensembles, MCS are based on the combination of different classifier algorithms and hence the individual advantages of each method can be combined. In terms of combination style, three categories can be applied: 1) Concatenation combination (the classification result generated by a classifier is used as the input into the next classifier until a result is obtained through the final classifier in the chain); 2) Parallel combination (multiple classifiers are designed independently and their outputs are combined according to certain strategies; and 3) Hierarchical combination (combines both concatenation and parallel combination) (Ranawana and Palade, 2006). According to the classifiers outputs, MCS in a parallel combination can be further divided into three levels: abstract level (each classifier outputs a unique label); rank level (classes are ranked based on decreasing likelihood); and measurement level (Based on probability values) (Ruta and Gabrys, 2000). But however, this review will be focused on the widely used parallel MCS combination. Some of the widely and successfully applied MCS approaches are:

9.2.1. Maximum Rule (MR) : MR is a simple method for combining probabilities provided by multiple classifiers. It interprets each class membership as a vote for one of the k classes. For each individual classifier, the class that receives the highest class membership is taken as the class label for that classifier. After that, the class labels from the N classifiers are compared again and the class that receives the highest class membership is taken as the final classification as in equation 12. is the class membership of a pixel to belong to a class Ck given by classifier fi, and PMR is the probability based on MR. The major problem of maximum rule is that all the classifiers have the same authority regardless of their reliability (Foody et al., 2007).

$$P_{MR} = \max\left[\max\ pp(C_k \mid f_i)\right] \qquad (12)$$

9.2.2. Weighted Sum (WS) : First, the class membership at the output of each classifier is weighted according to the classifiers' reliability (accuracy) for each class ($0 \leq \alpha ci \leq 1$). After that and for each class, the class memberships at the output of all classifiers are summed together. Finally, the class that receives the maximum sum is taken as the final class label as in equation 13. PWS is the probability based on the weighted sum, αci is the weight of each classifier, ppi: is the class membership value obtained for the ith classifier and N is the Number of classifiers (Le et al., 2007).

$$P_{WS} = \sum_{i=1}^{N} \alpha_{ci}\ pp_i \qquad (13)$$

9.2.3. Fuzzy Majority Voting (FMV) : The idea is to give some semantics or meaning to the weights so that the values for the weights can be provided directly. In the following, the semantics based on fuzzy linguistic quantifiers for the weights are used (Zadeh, 1983). First, the membership function of relative quantifiers can be defined as in equation 14 (Herrera and Verdegay, 1996). The parameters a, b [0, 1] and ppi is the class membership of pixel i. Then, Yager (1988) proposed to compute the weights based on the linguistic quantifier represented by Q as in equation 15.

$$Q_{P_i} = \begin{cases} 0 & if\ \ pp_i < a \\ \dfrac{pp_i - a}{b - a} & if\ \ a \leq pp_i \leq b \\ 1 & if\ \ pp_i > b \end{cases} \qquad (14)$$

$$w_{P_i} = Q_{P_i}\left(\frac{i}{N}\right) - Q_{P_i}\left(\frac{i-1}{N}\right),\ for\ i = 1, ..., N \qquad (15)$$

Q_{P_i} is the membership functions of relative quantifiers, i is the order of the classifier after ranking for all classifiers in a descending order and N is the total number of classifiers. A relative quantifier 'at least half' with the parameter pair (0, 0.5) is normally applied for the membership function Q in equation 14. Then, depending on the total number of classifiers N, and from equation 15 the corresponding weighting vector W = [w1, ….., wR] can be obtained. The final combined probability can be calculated as in equation 16, with is the weight based on linguistic quantifier, ppi is the Markovian probability of pixel i and k is the number of classes.

$$P_{FMV} = \arg\max_{k}\left[\sum_{i=1}^{N} w_{p_i}\ pp_i\right] \qquad (16)$$

9.2.4. Dempster-Shafer Theory (DST) : The theory of evidence was introduced by Shafer (1976) for combination of different measures of evidence. It is a generalization of the Bayesian framework and permits the characterization of uncertainty and ignorance. Consider a classification problem where the input data are to be classified into n classes Cj ∈ θ, θ is referred to as the frame of discernment. The power set of θ is denoted by 2θ (the set of all subsets of θ). A probability mass m (A) is assigned to every class A∈2θ by a classifier such that m (Ø) = 0, $0 \leq$ m (A) ≤ 1, and Σ m

(A) =1, and Ø denotes the empty set. m (A) can be interpreted as the amount of belief that is assigned exactly to A and not to any of its subsets. Imprecision of knowledge can be handled by assigning a non-zero probability mass to the union of two or more classes Cj. The support Sup (A) of a class A∈2θ is the sum of all masses assigned to that class. The plausibility Pls (A) sums up all probability masses not assigned to the complementary hypothesis Ā of A with A∩ Ā =Ø and A Ā = θ:

$$Sup(A) = \sum_{B \subseteq A} m(B); \qquad Pls(A) = \sum_{A \cap B \neq \phi} m(B) = 1 - Sup(\bar{A}) \tag{17}$$

Sup (A) is also called dubiety. It represents the degree to which the evidence contradicts a proposition. If k classes are available, probability masses mi (Bj) have to be defined for all these classes i with $1 \leq i \leq z$ and Bj 2θ.

From these probability masses, a combined probability mass can be computed for each class A 2θ as follow:

$$m(A) = \frac{\sum_{B_1 \cap B_2 \cap \ldots . B_z = A} \left[\prod_{1 \leq i \leq z} m_i(B_j) \right]}{1 - \sum_{B_1 \cap B_2 \cap \ldots . B_z = \phi} \left[\prod_{1 \leq i \leq z} m_i(B_j) \right]} \tag{18}$$

As soon as the combined probability masses m (A) are determined, both Sup (A) and Pls (A) can be computed. The accepted hypothesis Ca ∈ θ is determined according to a decision rule (the class of maximum plausibility or the class of maximum support). It is worth mentioning that the combination rule given by equation 18 assumes that the belief functions to be combined are independent. Many researchers have compared MCS. Table 10 provides summary of different researchers' conclusion and the situation in which each MCS is most useful.

Table 10: Performance evaluation of different MCSs

Researcher	Classifier	Datasets	MCS	Best Performance
Ebeir et al. (2001)	- ANN - CTs	- HR satellite imagery. - Spectral, spatial and contextual information.	- Bagging	- Bagging
Briem et al. (2002)	- MD - MXL - CTs	- SAR data. - Topographical data.	- Bagging - Bagging	- Boosting
Kumar et al. (2002)	- MXL	- Hyperspectral data	- Hierarchical fusion	- Hierarchical fusion
Waske and Benediktsson (2007)	- SVMs	- SAR data. - multispectral imagery	- CE/SVM	- CE/SVM
Ceamanos et al. (2010)	- SVMs	- Hyperspectral data	- CE/SVM	- CE/SVM
Trinder et al. (2010)	- SVMs - SOM - CTs	- Aerial Images - LiDAR data	- MR - WS - FMV - DST	- DST
Du et al. (2012)	- MLP - CTs - MD - SVMs - SAM - ArtMap	- QuickBird - OMISII - Landsat ETM+	- BPT - FMV - DST - CE/SVM	- BPT
	- MLP (Base classifier) - CTs	- QuickBird - OMISII	- Bagging - Boosting	- Boosting
	- MLP - SVMs - ArtMap - CTs	- QuickBird - Landsat ETM+	- MR - WS - FMV - DST -	- FMV
Ko et al. (2014)	- RFS - KNN - SVMs	- LiDAR data	- average voting	- average voting
Salah (2014)	- PB SVMs - OO SVMs	- IKONOS	- BPT	- BPT

10. Post classification processing

Post classification techniques can eliminate the short comings associated with classification algorithms such as unclassified or misclassified pixels, and hence improve the classification accuracy (Lu and Weng, 2007). The commonly used post classification techniques include: majority filter (MF); probability label relaxation (PLR); and cellular automata (CA) (Espinola et al., 2008). The MF reclassifies the center pixel when it is not a member of the majority class. It improves the overall accuracy of classification but merges some land cover classes together. The PLR is an iterative technique which considers the probabilities of the neighboring pixels for updating the probability of the center pixel. The PLR technique provides higher accuracy than the MF method, but it requires lot of computation. The approach of CA consists of regular grid of cells. Each cell is associated with a particular state from a set of possible states. The CA reassigns a class of the pixel according to the class of the neighboring pixels and based on a set of defined rules. In terms of accuracy, the CA approach has proved to be better than other two filters (Minu and Bindhu, 2016). On the other hand, ancillary data can be integrated after image classification. This can be done through very specific strategies such as: expert systems, rule based systems; and knowledge base systems.

11. Classification of accuracy assessment

Many sources of errors can affect the classification results which include: classification error, error from registration, and poor quality of training (Powell et al., 2004). These errors generate uncertainties (where is the error?) at different stages in the classification process which may influence the classification accuracy, as well as the estimated area of land-cover classes Posterior probabilities are an indicator of the uncertainty in making a particular class allocation. Accuracy assessment allows analyst to evaluate the utility of the resulted thematic map for the intended applications. In order to assess the classification accuracy, the classification results can be compared against the reference data. DeFries and Chan (2000) suggested the use of multiple criteria to evaluate the performance of algorithms. These criteria include classification accuracy, computational resources, stability, and robustness to noise in the training data. Classification accuracy is the most important criteria to evaluate the classification performance. The most common used methods for accuracy assessment are:

11.1 Overall Classification Accuracy

The overall accuracy is the most widely used approach for the evaluation of the classification results and can be calculated by equation 19:

$$\mathrm{OCA} = \frac{\mathrm{NCP}}{\mathrm{NRP}} \qquad (19)$$

Where OCA is the overall classification accuracy; NCP is the total number of correctly classified pixels (along the diagonal of the error matrix) and NRP is the total number of reference pixels. The error matrix is a simple cross tabulation of the resulted class label against the observed one in the reference data. Since the OCA is a global measure the performance of the classifier should also be evaluated by determining some other criteria as shown below.

11.2 Kappa Index of Agreement (KIA)

The Kappa Index of Agreement (KIA) is a statistical measure adapted for accuracy assessment in remote sensing fields by Congalton and Mead (1983). KIA tests two images, if their differences are due to chance or real disagreement. It is often used to check for accuracy of classified satellite images versus some real ground-truth data as in equation 20. For the per-category-Kappa, equation 21 was introduced by Rosenfield and Fitzpatrick-Lins (1986):

$$k = \frac{N\sum_{i=1}^{r} X_{ii} - \sum_{i=1}^{r}(X_{i+} * X_{+i})}{N^2 - \sum_{i=1}^{r}(X_{i+} * X_{+i})} \qquad (20)$$

r: number of row in the error matrix.
xii : number of combinations along the diagonal.
xi+ : total observations in row i.
x+i : total observations in column i.
N : total number of cells.

$$k_i = \frac{P_{ii} - P_{i+}P_{+i}}{P_{i+} - P_{i+}P_{+i}} \qquad (21)$$

pii : proportion of units agreeing in row i / column i
pi+ : proportion of units for expected chance agreement in row i
p+i : proportion of units for expected chance agreement in column i

11.3 Omission and Commission Errors

Unlike OCA, commission and omission errors clearly show whether the proposed classifier improves or deteriorates the results for each individual class compared to the reference data (Congalton, 1991).

$$CE_I = \frac{A_1 + A_2 + A_3}{R_1} \qquad (22)$$

$$OE_I = \frac{B_1 + B_2 + B_3}{R_1} \qquad (23)$$

CEI and OEI are commission and omission errors of class increased; A1, A2 and A3 are the numbers of incorrectly identified pixels of class increased associated with classes decreased, background and unchanged; R1 is the total number of pixels of the class increased as observed in the reference data; B1, B2 and B3 are the numbers of unrecognized pixels that should have identified as belonging to the class increased. The same is applicable for the class decreased.

12. Commercial software

The availability of classification software is one of the most important factors that must be taken into account when selecting a classification method for use. Various image processing software packages make it possible to enhance, analyze, interpret and extract meaningful information from remotely sensed data. Table 11 lists the most common used image processing packages along with the available classification approaches. This table is intended to be highly useful for those wishing to select the most appropriate software for the problem under investigation.

Table 11: Classification techniques available in the commonly used commercial software

IDRISI	ENVI	Erdas Imagine	ILWIS
- ISODATA	- ISODATA	- ISODATA	- PP
- K-means	- K-means	- MD	- MD
- PP	- CTs	- MXL	- MhD
- MD	- SVM	- MhD	- MXL
- MhD	- PP	- Expert Classifier	
- MXL	- MD		
- Fisher LDA	- MhD		
- KNN	- MXL		
- CTs	- SAM		
- MLP	- RBF		
- SOM			
- Fuzzy ArtMap			
- RBF			
- Bayesian probability			
- Fuzzy set			
- Linear Spectral Unmixing			

13. Summary and discussion

The most suitable classification algorithm is based on the spatial resolution of the used satellite imagery. In the case of HR data such as IKONOS, SPOT 5 HRG and World View-2, per-field and object-oriented classifiers may outperform the per-pixel ones. On the other hand, the integration of spectral and texture information can reduce the problem of shadow and the wide spectral variation within the land-cover classes. In the case of medium and coarse spatial resolution, sub-pixel classifiers have proved to be more useful than per-pixel classifiers because of the mixed pixels problem. In this case, the loss of spatial information makes spectral information more important than spatial one. Furthermore, ancillary data can be integrated with spectral data for improved classification results.

The optimum training sample size varies from one classifier to another. Selection of proper size of samples are important factors which governs the classification accuracy. All classifiers are shared in the same behavior of after certain size of training sample, the classification accuracy showed downward trend with the increasing size of training data. In the case of limited number of training samples, SVM and maximum likelihood have proved to be the best choice. When multisource data are used, parametric classifiers such as MXL are typically not appropriate for image classification. Advanced non-parametric classifiers, such as ANN, SVMs and CTs can be more suitable.

There are several ANN approaches that can be used to classify remotely sensed images which include: MLP; SOM; and Fuzzy ArtMap. Fuzzy ArtMap has proved to be the most efficient algorithms, followed by the MLP. SOM produced the lowest classification accuracy in the majority of articles. All these algorithms depend mainly on the operators experience in setting up their parameters in order to reach the optimal performance. MLP requires a complete retraining of the whole network. This may lead to long training time, even for small size test areas. Fuzzy ArtMap, on the other hand, can solve large scale problems through a few training epochs. The only defect with Fuzzy ArtMap is that it is sensitive to noise and outliers that may decrease the classification accuracy. Unlike MLP and Fuzzy ArtMap, SOM allows for the discrimination of multimodal classes. On the other hand, SOM normally yields many unclassified pixels.

In case of CTs, the Entropy splitting algorithm has proved to be a preferable algorithm for image classification. On the other hand, the 10-fold cross validation process has proved to be an accurate method. As well, CT derived from a given test area could be successfully transferred to another area provide the remotely sensed images having the same sensor characteristics and the LULC are similar. In general,

SVMs outperform other classifiers in terms of classification accuracy. SVMs show a balance between errors of the classes. In some cases, the RBF kernel would be the best choice. However, a grid search with a 10-fold cross validation has to be applied to search for the RBF kernel parameters, C and γ for the SVM classifier.

Different classifiers offer complementary information about the data to be classified. One classifier might be more efficient at detecting a specific class, while another classifier is more efficient for another specific one. Combining classifiers in an efficient way can improve classification accuracy than any single classifier, even the best one. Neural and statistical classifiers result in uncorrelated classification errors and hence higher classification accuracies can be obtained by combining them. It is worth mentioning that adding more classifiers to the system does not guarantee improvements in the performance. However, diversity is an important requirement for the success of hybrid systems. The combination selected by Disagreement accuracy measure usually outperforms the ones selected by other diversity measures. Two approaches exist for combining classifiers: 1) CE; and 2) MCS. Classifier ensembles are based on the combination of a variety of the same algorithm. On the other hand, MCS are based on the combination of different classification algorithms. Most of the existing MCSs suffer one or more shortcomings such as: high ambiguities between classes; high sensitivity to noise in the data; and high computational load. D-S combination, as a MCS, has proved to be superior to other hybrid systems in terms of classification accuracy.

14. Conclusion

Image classification has made great progress over the past few decades in the development and use of advanced classification algorithms. This review gives a brief guide about different classification techniques and lists the advantages and disadvantages of each. It is concentrated extensively on recent classification algorithms such as ANN, SVMs and CTs. These classification approaches have significantly improved the accuracy of the results in the case of HR satellite imagery. This paper helps researchers in selecting a suitable classification algorithm for a specific task, optimization of the classifiers and selecting the optimal classifiers for constructing MCS. Most of the MCS can enhance classification accuracy, but the performances are affected by different factors such as the selected base classifiers and the combination strategy. Diversity measures can play a vital rule in selecting the base classifiers for a MCS.

References

Abburu, S. and S. Golla (2015). Satellite image classification methods and techniques: A review. International Journal of Computer Applications, 119 (8): 20-25.

Aksela, M. and J. Laaksonen (2006). Using diversity of errors for selecting members of a committee classifier. Pattern Recognition, 2006(39): 608–623.

Al-doski, J., S. Mansor and H. Shafri (2013). Image classification in remote sensing. Journal of Environment and Earth Science, (3)10: 141-148.

Anthony, G., H. Gregg and M. Tshilidzi (2007). Image classification using SVMs: One-against-one Vs one-against-all. Proceedings of the 28th Asian Conference on Remote Sensing ARCS, Learning (cs.LG); Artificial Intelligence (cs.AI); Computer Vision and Pattern Recognition (cs.CV) , Kuala Lumpur, Malaysia, 12-16 November 2007.

Baban, S.M.J. and K.W. Yusof (2001). Mapping land use/cover distribution on a mountainous tropical island using remote sensing and GIS. International Journal of Remote Sensing, 22(10): 1909–1918.

Bauer, E. and R. Kohavi (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning, 36(1): 105–139.

Benediktsson, J.A., J. Chanussot and M. Fauvel (2007). Multiple classifier systems in remote sensing: From basics to recent developments. MCS 2007, LNCS 4472, (M. Haindl, J. Kittler, and F. Roli, editors), Springer Verlag, Berlin 2007: 501-512.

Bezdec, J.C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2):123–140.

Breiman, L. (2001). Random forest. Machine Learning, 45(1): 5–32.

Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone (Ed.) (1984). Classification and regression trees. 358 p (New York: Chapman & Hall).

Briem, G., J. Benediktsson and J. Sveinsson (2002). Multiple classifiers applied to multisource remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, 40 (10): 2291–2299.

Bronge, L.B. (1999). Mapping boreal vegetation using Landsat TM and topographic map data in a stratified approach. Canadian Journal of Remote Sensing, 25(5): 460–474.

Caetano, M. (2009). Image classification. An ESA Advanced Training Course on Land Remote Sensing, 28 June-03 July 2009 in Prague, Czech Republic.

Camps-Valls, G., L. Gomez-Chova, J. Calpe-Maravilla, E. Soria-Olivas, J.D. Martin Guerrero and J. Moreno (2003). Support vector machines for crop classification

using hyperspectral data. Proceedings of ibPRIA, Mallorca, Spain, 4-6 June 2003: 134-141.

Carpenter G.A., S. Crossberg and J.H. Reynolds (1991). ARTMAP: Supervised real time learning and classification of nonstationary data by a self-organizing neural network. Neural Networks, 4(5): 565-588.

Ceamanos, X., B. Waske, J.A. Benediktsson, J. Chanussot, M. Fauvel and J.R. Sveinsson (2010). A classifier ensemble based on fusion of support vector machines for classifying hyperspectral data. International Journal of Image and Data Fusion, 1 (4): 293–307.

Chaichoke, V., P. Supawee, V. Tanasak and K.S. Andrew (2011). A Normalized Difference Vegetation Index (NDVI) time-series of idle agriculture lands: A preliminary study. Engineering Journal, 15(1): 9-16.

Chandra, A. and X. Yao (2006). Evolving hybrid ensembles of learning machines for better generalisation. Neurocomputing, 69(7-9): 686–700.

Collins, M., R.E. Schapire and Y. Singer (2002). Logistic regression, Adaboost and Bregman distances. Machine Learning, 48(1): 31–44.

Congalton, R.G. and R.A. Mead (1983). A quantitative method to test for consistency and correctness in photointerpretation. Photogrammetric Engineering and Remote Sensing, 49(1): 69 – 74.

Congalton, R.G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment, 37(1): 35–46.

Cybenko, G. (1989) Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems, 2(4): 303–314.

Defries, R.S. and J.C. Chan (2000). Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. Remote Sensing of Environment, 74(3):503–515.

Doma, M.L., M.S. Gomaa and R.A. Amer, R.A. (2015) Sensitivity of pixel-based classifiers to training sample size in case of high resolution satellite imagery. Journal of Geomatics, 9(2): 53-58.

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

Eastman, J.R. (2006). Idrisi Andes: Tutorial. Clark Labs. Clark University, Worcester.

Ebeir, L.D., P.A.L. Atinne and I.S. Teen (2001). Remote sensing classification of spectral, spatial and contextual data using multiple classifier systems. Proceedings of the 8th ECS and Image Analysis, September 4-7, Bordeaux, France, 584–589.

Epstein, J., K. Payne and E. Kramer (2002). Techniques for mapping suburban sprawl. Photogrammetric Engineering and Remote Sensing, 63(9): 913–918.

Espinola, M., R. Ayala, S. Leguizamon and M. Menenti (2008). Classification of satellite images using the cellular automata approach. Proceedings of the 1st WSKS, CCIS, 19: 521-526.

Foody, G.M. (1995). Land-cover classification by an artificial neural network with ancillary information. International Journal of Geographical Information Systems, 9(5): 527–542.

Foody, G.M. (1999). Image classification with a neural network: From completely crisp to fully-fuzzy situations. In P.M. Atkinson and N.J. Tate (eds), Advances in Remote Sensing and GIS analysis, Chichester: Wiley&Son.

Foody, G.M. (2009). Classification accuracy comparison: Hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. Remote Sensing of Environment, 113(8), 1658–1663.

Foody, G.M., D.S. Boyd and C. Sanchez-Hernandez (2007). Mapping a specific class with an ensemble of classifiers. International Journal of Remote Sensing, 28(8): 1733–1746.

Freund, Y. and R.E. Schapire (1997). A decision-theoretic generalization of online learning and application to boosting. Journal of Computer and System Science, 55(1): 119-139.

Giacinto, G. and F. Roli (2001). Design of effective neural network ensembles for image classification. Image and Vision Computing, 19(9-10): 697–705.

Gil, A., Q. Yu, A. Lobo, P. Lourenço, L. Silva and H. Calado (2011). Assessing the effectiveness of high resolution satellite imagery for vegetation mapping in Small islands protected areas. Journal of Coastal Research, 64(2011): 1663-1667.

Gislason, P.O., J.A. Benediktsson and J.R. Sveinsson (2006). Random forests for land cover classification. Pattern Recognition Letters, 27(4): 294-300.

Groom, G.B., R.M. Fuller and A.R. Jones (1996). Contextual correction: Techniques for improving land cover mapping from remotely sensed images. International Journal of Remote Sensing, 17(1): 69–89.

Hadjimitsis, D.G., C.R.I. Clayton and V.S. Hope (2004). An assessment of the effectiveness of atmospheric correction algorithms through the remote sensing of some reservoirs. International Journal of Remote Sensing, 25(18): 3651–3674.

Hale, S.R. and B.N. Rock (2003). Impacts of topographic normalization on land-cover classification accuracy. Photogrammetric Engineering and Remote Sensing, 69(7): 785–792.

Hamedianfar, A., H.Z. Mohd Shafri, S. Mansor and N. Ahmad (2014). Detailed urban object-based classifications from WorldView-2 imagery and LiDAR data: Supervised vs. fuzzy rule-based. FIG Congress 2014, Engaging the Challenges—Enhancing the Relevance, Kuala Lumpur, 16-21 June 2014.

Helmer, E.H., S. Brown and W.B. Cohen (2000). Mapping montane tropical forest successional stage and land use with multi-date Landsat imagery. International Journal of Remote Sensing, 21(11): 2163–2183.

Herrera, F. and J.L. Verdegay (1996). A linguistic decision process in group decision making. Group Decision Negotiation, 5(2): 165-176.

Ho, T.K. (1998) The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8): 832–844.

Hsu, C.W., C.C. Chang and C.J. Lin (2009). A practical guide to support vector classification. Department of Computer Science, National Taiwan University, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (Accessed 7 September 2016).

Hugo, C., L. Capao, B. Fernando and C. Mario (2007). MERIS based land cover classification with self-organizing maps: Preliminary results. In Proceedings of the 2nd EARSeL SIG Workshop on Land Use & Land Cover (unpaginated CD-ROM), 28 – 30 September 2006, Bonn, Germany.

Hwang, Y.S. and S.Y. Bang (1997). An efficient method to construct a radial basis function neural network classifier. Neural Networks, 10(8): 1495-1503.

Jawak, S., P. Devliyal and A. Luis (2015). A comprehensive review on pixel oriented and object oriented methods for information extraction from remotely sensed satellite images with a special emphasis on cryospheric applications. Advances in Remote Sensing, 4(3): 177-195.

Jen-Hon, L. and T. Din-Chang (2000). Self-organizing feature map for multi-spectral spot land cover classification. GIS development.net, AARS, ACRS 2000.

Jensen, J. (2005). Introductory Digital Image Processing, Third Edition, Prentice Hall, 526 p.

Kamavisdar, P., S. Saluja and S. Agrawal (2013). A survey on image classification approaches and techniques. International Journal of Advanced Research in Computer and Communication Engineering, 2(1): 1005-1008.

Kanellopoulos, I., G. Wilkinson, F. Roli and J. Austin (editors) (1997). Neurocomputation in remote sensing data analysis. Springer, Berlin.

Kavzoglu, T. and P.M. Mather (2003). The use of back propagating artificial neural networks in land cover classification. International Journal of Remote Sensing, 24(3): 4907- 4938.

Ko, C., G. Sohn, T. Remmel and J. Miller (2014). Hybrid ensemble classification of tree genera using airborne LiDAR data. Remote Sensing, 6 (x): 11225-11243.

Kohonen, T. (1990) The self-organizing map. Proceedings of the IEEE, 78: 1464-80.

Kumar, M. and R.K. Singh (2013). Digital image processing of remotely sensed satellite images for information extraction. Conference on Advances in Communication and Control Systems (CAC2S 2013), Atlantis Press, pp. 406-410.

Kumar, S., J. Ghosh and M.M. Crawford (2002). Hierarchical fusion of multiple classifiers for hyperspectral data analysis. Pattern Analysis and Applications, 5: 210–220.

Kuncheva, L.I. and C.J. Whitaker (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning, 51(2): 181–207.

Kussul, N., S. Skakun and O. Kussul (2006). Comparative analysis of neural networks and statistical approaches to remote sensing image classification. Computing, 5(2): 93-99.

Le, A.C., V.N. Huynh, A. Shimazu and Y. Nakamori (2007). Combining classifiers for word sense disambiguation based on Dempster-Shafer theory and OWA operators. Data and Knowledge Engineering, 63 (2): 381-396.

Lefsky, M.A. and W.B. Cohen (2003). Selection of remotely sensed data. In M.A. Wulder and S.E. Franklin (Eds), Remote Sensing of Forest Environments: Concepts and case studies, 13– 46 (Boston: Kluwer Academic Publishers).

Li, G., D. Lu, E. Moran and S. Sant'Anna (2012). Comparative analysis of classification algorithms and multiple sensor data for land use/land cover classification in the Brazilian Amazon. Journal of Applied Remote Sensing 6(1): 11 pages.

Lillesand, T. and R. Kiefer (2004). Remote Sensing and Image Interpretation. Fourth Edition, John Willey & Sons, Inc., New York.

Lippitt, C., J. Rogan, Z. Li, J. Eastman and T. Jones (2008). Mapping selective logging in mixed deciduous forest: A comparison of machine learning algorithms.

Photogrammetric Engineering and Remote Sensing, 74(10): 1201–1211.

Liu, W., K. Seto, E. Wu, S. Gopal and C. Woodcock (2004). ARTMMAP: A neural network approach to subpixel classification. IEEE Transactions on Geoscience and Remote Sensing, 42(9): 1976–1983.

Mannan B., J. Roy and A.K. Ray (1998). Fuzzy ArtMap supervised classification of multi-spectral remotely-sensed images. International Journal of Remote Sensing, 19(4): 767-774.

Maryam, N., M.Z. Vahid and H. Mehdi (2014). Comparing different classifications of satellite imagery in forest mapping (Case Study: Zagros Forests in Iran). International Research Journal of Applied and Basic Sciences, 8(7): 1407-1415.

Maselli, F., A. Rodolfi, L. Bottai, S. Romanelli and C. Conese (2000). Classification of Mediterranean vegetation by TM and ancillary data for the evaluation of fire risk. International Journal of Remote Sensing, 21(17): 3303–3313.

Michail, P., J.A. Benediktsson and K. Ioannis (2002). The effect of classifier agreement on the accuracy of the combined classifier in decision level fusion. IEEE Transactions on Geoscience and Remote Sensing, 39(11): 2539–2546.

Minu, N.S. and J.S. Bindhu (2016). Supervised techniques and approaches for satellite image classification. International Journal of Computer Applications, 134(16): 0975 – 8887.

Mountrakis, G., J. Im and C. Ogole (2011). Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 66(3): 247-259.

Nasrabadi, N.M. and Y. Feng (1988). Vector quantization of images based upon the Kohonen self-organizing feature maps. Proceedings of the IEEE International Conference on Neural Networks (ICNN-88), 24-27 July 1988, San Diego, California, 101-108.

Oliveira, L., T. Oliveira, L. Carvalho, W. Lacerda, S. Campos and A. Martinhago (2007). Comparison of machine learning algorithms for mapping the phytophysiognomies of the Brazilian Cerrado. IX Brazilian Symposium on GeoInformatics, Campos do Jordão, Brazil, November 25-28, 2007, INPE, 195-205.

Pal, M. and P. Mather (2005). Support vector machines for classification in remote sensing. International Journal of Remote Sensing, 26(5): 1007-1011.

Powell, R.L., N. Matzke, C. De Souza Jr, M. Clark, I. Numata, L.L. Hess and D.A. Roberts (2004). Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. Remote Sensing of Environment, 90(2): 221–234.

Prasad, S., T. Savithri and I. Murali Krishna (2015). Techniques in image classification; A survey. Global Journal of Researches in Engineering: Electrical and electronics Engineering, 16(6): 17-32.

Qiu, F. and J.R. Jensen (2004). Opening the black box of neutral networks for remote sensing image classification. International Journal of Remote Sensing, 25(9): 1749–1768.

Quinlan, J.R. (1987). Simplifying decision trees. International Journal of Man-Machine Studies, 27(3): 227–248.

Ranawana, R. and V. Palade (2006). Multi-classifier systems: Review and a roadmap for developers. International Journal of Hybrid Intelligent Systems, 3(1): 35–61.

Richards, J.A. (2013). Remote sensing digital image analysis. Springer-Verlag, Berlin, 5th Ed. 496 p.

Rodriguez, J.J. and L.I. Kuncheva (2009). Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(10): 1619–1630.

Rosenblatt, F. (1962). Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. Spartan Books, Washington DC, 1962.

Rosenfield, G.H. and K. Fitzpatrick-Lins (1986). A coefficient of agreement as a measure of thematic classification accuracy. Photogrammetric Engineering and Remote Sensing, 52(2): 223 – 227.

Ruta, D. and B. Gabrys (2007). An overview of classifier fusion methods. Computing and Information Systems, 2000(7): 1–10.

Salah, M. (2014). Combining pixel-based and object-oriented support vector machines using Bayesian probability theory. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume II-7, 2014 ISPRS Technical Commission VII Symposium, 29 September – 2 October 2014, Istanbul, Turkey.

Shafer, G. (1976). A mathematical theory of evidence. Princeton University Press.

Shaker, A., W.Y. Yan and N. El-Ashmawy (2012). Panchromatic satellite image classification for flood hazard assessment. Journal of Applied Research and Technology, 10 (x): 902-911.

Shannon, C.E. (Ed.) (1949). The mathematical theory of communication. (Urbana, IL: University of Illinois Press).

Sherrod, P.H. (2008). DTREG tutorial home page. Available online at:

http://www.dtreg.com/crossvalidation.htm (Accessed 7 September 2016).

Stefanov, W.L., M.S. Ramsey and P.R. Christensen (2001). Monitoring urban land cover change: An expert system approach to land cover classification of semiarid to arid urban centers. Remote Sensing of Environment, 77(2): 173–185.

Trinder, J., M. Salah, A. Shaker, M. Hamed and A. Elsagheer (2010). Combining statistical and neural classifiers using Dempster-Shafer theory of evidence for improved building detection. 15th ARSPC, Alice Springs, Australia, 13- 17 September 2010.

Tso, B. and P.M. Mather (2009). Classification methods for remotely sensed data. 2nd Ed. Chapter 2-3, Taylor and Francis Group, America.

Tso, B.C.K. and P.M. Mather (1999). Classification of multisource remote sensing imagery using a genetic algorithm and Markov random fields. IEEE Transactions on Geoscience and Remote Sensing, 37(3): 1255–1260.

Van der Linden, S., A. Rabe, A. Okujeni and P. Hostert (2009). Image SVM classification. application manual: imageSVM version 2, Humboldt-Universität zu Berlin, Germany.

Vapnik, V. (1979). Estimation of dependences based on empirical data [in Russian]. Nauka, Moscow, 1979. (English translation: Springer Verlag, New York, 1982).

Vesanto, J., J. Himberg, E. Alhoniemi and J. Parhankangas (2000). SOM toolbox for Matlab. Technical Report A57, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland.

Waske, B. (2007). Classifying multisensor remote sensing data: Concepts, algorithms and applications. PhD thesis, Bonn University, Germany.

Waske, B. and J.A. Benediktsson (2007). Fusion of support vector machines for classification of multisensory data. IEEE Transactions on Geoscience and Remote Sensing, 45(12): 3858–3866.

Webb, G.I. (2009). Multiboosting: A technique for combining boosting and wagging. Machine Learning, 40(2): 159–196.

Wilkinson, G.G. (2005). Results and implications of a study of fifteen years of satellite image classification experiments. IEEE Transaction on Geosciences and Remote Sensing, 43(3): 433-440.

Yager, R.R. (1988). On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Transactions on Systems, Man, and Cybernetics, 18(1): 183-190.

Yu-Chang, T. and C. Kun-Shan (2009). An adaptive thresholding multiple classifiers system for remote sensing image classification. Photogrammetry Engineering and Remote Sensing, 75(6): 679-687.

Zadeh, L.A. (1983). A computational approach to fuzzy quantifiers in natural languages. Computers and Mathematics with Applications, 9(1): 149-184.

Zhang, C., and J. Zhang (2008). RotBoost: A technique for combining rotation forest and AdaBoost. Pattern Recognition Letters, 29(10): 1524–1536.

Zhang, Q., J. Wang, X. Peng, P. Gong and P. Shi (2002). Urban built-up land change detection with road density and spectral information from multitemporal Landsat TM data. International Journal of Remote Sensing, 23(15): 3057–3078.

Zhang, Y. (1999). Optimization of building detection in satellite images by combining multispectral classification and texture filtering. ISPRS Journal of Photogrammetry and Remote Sensing, 54(1): 50– 60.