# MODULE – 5 LECTURE NOTES – 1

# IMAGE CLASSIFICATION - SUPERVISED

## 1. Introduction

Classification is the technique by which real world objects/land covers are identified within remotely sensed imagery. Consider a multispectral image with m bands. In the case of a simplest pixel, its characteristics are expressed in the form of a vector where the vector elements represent the spectral properties of that pixel, which can be captured in these m bands. The number of classes can be determined a priori or using certain indices. The classes represented by this pixel may be water bodies, woodlands, grassland, agriculture, urban or other land cover types. Classification identifies the land cover represented by each pixel based on its spectral reflectance value (or digital number). The process also involves labeling each class entities using numerical value which can be done using a classification rule or a decision rule. In this context, the process of clustering involves an exploratory procedure wherein the aim is to estimate the number of distinct land cover classes present in an area and also to allocate pixels to these aforementioned classes. Image classification can be of many types. The two major classification types are supervised and unsupervised. These two techniques of pixel labeling can also be utilized to segment an image into regions of similar attributes. Features/ patterns can be defined using the spectral information present in bands. In other words, a pattern associated with each pixel position within an image is identified. Pattern recognition methods have been widely applied in various fields of engineering and sciences. This lecture will explain the supervised classification technique. Please note that for better understandability regarding some of the terms discussed in this module, the readers are expected to have minimum knowledge regarding basic statistics.

## 2. Supervised Classification

In supervised classification technique, the location of land cover types should be known a priori. The areas of each land cover types are known as training sites. The spectral characteristics of pixel digital numbers within each of the land cover types can be used to generate multivariate

statistical parameters for each of the training sites. As the supervised classification methods are based on statistical concepts, this classification is also termed as per-point or per-pixel classification. One of the earlier methods adopted to visualize the distribution of spectral values measured on two features (for example, water body and agriculture land) was to generate a scatter plot. Visual inspection will reveal the existence of two separate land use types. This sheds light on two fundamental ideas of classification. First is using Euclidean space to represent the selected features of interest. And second is the usage of distance measure to club or estimate resemblance of pairs of points as a decision rule in order to classify the pixels as water body and agriculture land. Visual interpretation is intuitive and simple in nature. Eye and the brain together recognize the existence of two clusters or regions of feature space having a tight distribution of points with a relatively empty region in between them. We can even come up with a boundary line separating the two clusters which is called the decision boundary. This concept can as well be extended to three dimensions. Now the distance between clouds of points can be calculated to arrive at a decision boundary which will be a plane within a 3 dimensional feature space. Supervised classifiers require that the number of classes be specified in advance and that certain statistical characteristics of each class be known prior. This requires appropriate selection of training samples which is discussed in the next section.

### 3. Selection of training samples

Once a classification type is finalized, user is bound to select training sites within the imagery representative of various land cover types. Statistical classifiers depend on a good quality of training data to generate excellent results. Classification is done to collect spectral statistics for each of the land cover types. The quantity and quality of training data should be carefully chosen. Some people rely on a general thumb rule to select > 10m pixels for each class where m is the number of bands. A standard approach followed is to assume multinomial distribution of population and estimate the optimum number of pixels using the expression:

$$n = \frac{B.p_i.(1 - p_i)}{b_i^2}$$

Where, $p_i$ is the a priori probability for the class *i*. It is taken as half if we have no information about the area because at $p_i = \frac{1}{2}$ , *n* will be maximum. $b_i$ is the required absolute precision, *B* is defined as the upper $\frac{\alpha}{k} \times 100^{th}$ percentile for $\chi^2$ distribution with one degree of freedom, where *k* is the number of classes, and $\alpha$ is the desired confidence interval (Congalton and Green, 1998).

There exists a number of ways to select the training data. These include a) collecting in situ information b) selecting on screen polygonal training data and c) seeding of training data on screen. Most of the image processing softwares employ a polygonal tool which enables selection of region of interest (ROI). This can be used to select pixels corresponding to each of the land cover types as observed on screen. The users may also select a specific location (x,y) within the image using cursor. The seed program can then be used to evaluate and expand to neighboring pixels like an amoebae till it stops finding pixels with similar characteristics as the one originally selected. Both these methods are highly effective in training area selection. The final aim of training class selection is to achieve nonautoccorrelated training data as training data collected from autocorrelated data tends to possess reduced variance. An ideal approach would be to collect training data within a region using nth pixel based on some sampling technique. A common approach to conduct random sampling is to overlay the classified data over a grid so that cells within the grid can be selected randomly and groups of pixels within the test cells can be evaluated. After training data have been selected for each land cover type, the various statistical measures like mean, standard deviation, variance, minimum value, maximum value etc can be calculated and used for further classification processes. After selection of training data for each land cover type, it is essential to judge the importance of bands that will be highly effective in demarcating each class from all others. This process is known as feature selection which enables deletion of certain bands that provide redundant spectral information. Graphical methods of feature selection employ the usage of bar graph spectral plots, ellipse plots, feature space plots etc which are simple in nature and provide effective visual presentation of inter-class separability. These plots are sufficient to estimate inter class separability of 2 or 3 classes. But it would not be suitable to decide if a certain combination of four or five bands can perform better or not. For these, a statistical method of feature selection needs to be followed, which is discussed in the next section.

## 4. Statistical Feature Selection Measures

Spectral pattern recognition involving data values captured in m different bands needs to be subjected to suitable discrimination techniques in order that various land cover classes be separated with a minimum of error. Greater the number of bands available, greater will be the associated cost. Usually in the case of an overlap, either the pixel is assigned to a class to which it does not belong in the first place (commission error) or it is not assigned to its actual class (omission error). Statistical measures of feature selection provide measures to statistically separate the classes using the training data. Several measures exist, some of which are discussed below.

a) Divergence

Divergence is widely used a statistical measure of separability in machine processing of remotely sensed data. It basically suggests the best combination of subset bands (say q number) from a total of m bands that should be used during a supervised classification process. The number of combination of m bands taken q at a time is given as:

$$C_q^m = \frac{m!}{q!(m-q)!}$$

For example, suppose there are 6 bands of a multispectral imagery and are interested to know the three best bands to use it would be necessary to evaluate 20 combinations.

$$C_3^6 = \frac{6!}{3!(6-3)!} = 20 \text{ combinations}$$

Using training data of a supervised classification, the degree of divergence or separability between two classes (say a and b) can be computed using the expression:

$$Diver_{ab} = \frac{1}{2} Tr\left[(V_a - V_b)(V_b^{-1} - V_a^{-1})\right] + \frac{1}{2} Tr\left[(V_a^{-1} + V_b^{-1})(M_a - M_b)(M_a - M_b)^T\right]$$

Here $Tr$ = Trace of the matrix (Sum of diagonal elements)

$V_a, V_b$ = Covariance matrices of the two classes $a, b$

$M_a, M_b$ = Mean vectors of the two classes $a, b$

This expression enables one to identify the statistical separability between two classes using the given m bands of the training data. Now consider a situation with more than two classes. In such situation ideally the average over all possible pairs of classes needs to be computed holding the subset of bands (q) as constant. Then, another subset of bands are selected from these m classes to be then again analyzed. That subset which gives maximum divergence value needs to be selected as the superior set of bands to be further used in the classification algorithm.

$$Diver_{average} = \frac{\sum_{a=1}^{m-1} \sum_{b=a+1}^{m} Diver_{ab}}{C}$$

This has a drawback that outlying easily separable classes will mislead divergence which may result in the choice of suboptimal reduced feature subset as best. To compensate it becomes essential to compute the transformed divergence as:

$$TDiver_{ab} = 2000\left[1 - \exp\left(\frac{-Diver_{ab}}{8}\right)\right]$$

This measure scales the divergence values to lie in between 0 and 2000 providing an exponentially deceasing weight to increasing distances between various classes. As a result, a transformed divergence value of 2000 indicates excellent class separation.

b) Bhattacharyya Distance measure

This measure separates two classes at a time assuming that both the classes are Gaussian in nature and that their means and covariance matrices are available. The expression is given as:

$$Bhatt_{ab} = \frac{1}{8}(M_a - M_b)' \frac{(V_a - V_b)}{2}(M_a - M_b) + \frac{1}{2}\log_e \frac{\det \frac{V_a + V_b}{2}}{\sqrt{\det(V_a)}\sqrt{\det(V_b)}}$$

Here, det stands for the determinant of the matrix considered.

c) Jeffreys-Matusita Distance (JM Distance)

This measure provides a saturating behavior with increasing class separation similar to transformed divergence. But this measure is not as computationally effective as transformed divergence measure. The expression is given as:

$$JM_{ab} = \sqrt{2(1 - e^{-Bhatt_{ab}})}$$

**5. Parallelepiped classifier**

A parallelepiped refers to a prism whose bases are essentially parallelograms in nature. This classifier requires minimal user information in order to classify pixels. Assume a multispectral image with m bands or features that represents a total of n classes. For each of these n classes, the user is asked to provide the maximum and minimum values of pixels on each of these m bands. This allows a range to be created that can be expressed as given number of standard deviation units on either side of the mean of each feature. These enable setting boundary of the parallelepipeds that can be used/drawn to define regions within the m dimensional feature space. Hence, the decision rule used is relatively simpler wherein each pixel that is to be classified is taken one by one and determined if its value on the m bands lies inside or outside any of the parallelepipeds. Some pixels can be seen to be not lying inside any of the parallelepipeds. These can be classified as unknown category. In an extreme case scenario, overlapping parallelepipeds can be found enclosing same pixels. Decision making in such cases involves allocation of pixel to the first parallelepiped inside whose boundary it falls.

This classification technique is simple, quick which gives good results provided the data are well structured in nature. However, in practical scenario, this is rarely the case. When the image data sets in various bands result in overlapping parallelepipeds, a sensible decision rule is to calculate the Euclidean distance between the doubtful pixel and the centerpoint on each of the overlapping parallelepipeds. Then, a minimum distance rule can be employed to best decide on the output. In

other words, a boundary line is drawn in the area between two overlapping parallelepipeds. This boundary will be equidistant from the center points of the parallelepipeds and then pixels are classified based on their distance relative to the drawn boundary line. This technique is easy, fast in implementation. But it suffers from major drawbacks. The use of just minimum and maximum pixel values might not be a very good representative of the actual spectral classes present within an imagery. Also, it is assumed that the shape enclosing a particular spectral class can be neatly fit inside a parallelepiped which may not be necessarily so. Hence, this method is considered as a not so accurate a representation of feature space.

**6. Minimum distance to means classification algorithm**

As the name suggests, this classification technique utilizes a distance based decision rule. It requires the user to provide the mean vectors for each class in each band from the training data. Euclidean distance based on the Pythagorean theorem is normally employed to calculate the distance measure. For example, to calculate the Euclidean distance from a point (20,20) to the mean of class i (16,16) measured in bands 2 and 3 uses the expression

$$Dist = \sqrt{(BV_{ijk} - \mu_{ak})^2 + (BV_{ijl} - \mu_{al})^2}$$

Here $\mu_{ak}, \mu_{al}$ represents the mean vectors for class $a$ that has been captured in bands $k, l$ respectively.

Every pixel is assigned to a class based on its distance from the mean of each class. Many minimum distance algorithms specify a distance threshold from the class mean beyond which a pixel will not be assigned to that particular class even if it is nearest to the mean. This classification type is commonly used and computationally simple in nature.

## 7. Maximum likelihood classifier

Maximum likelihood classification relies on the assumption that geometrically; the shape of a cloud of points representing a particular class can be well represented using an ellipsoid. The orientation of the ellipsoid will depend on the degree of covariance among the features with an upward sloping major axis towards left indicating negative covariance, an upward sloping major axis toward right indicating high positive covariance and a near circular ellipse (with major axis ~ minor axis) indicative of lower covariances between the features. The statistical descriptors of mean, variance and covariance of features can be used to define the location, shape and size of the ellipse. We can consider a set of concentric ellipses each representing contours of probability of membership of the class with the probability of membership decreasing away from the centroid more rapidly along the minor axis than the major axis. Thus, a pixel is assigned to that class for which it has the highest probability of membership. This results in classification which is more accurate than the output by parallelepiped or k mean classification. As the training data are relied upon to produce shape of the distribution of the membership of each class. This classification assumes that the frequency distribution of class membership can be approximated using a multivariate normal probability distribution. Though the assumption of normality holds reasonable well, this does not work well where there are small departures from normality.

Assume X represents a vector of pixel values captured in m bands. In order to estimate if it belongs to class i, the probability of the pixel vector can be calculated using the multivariate normal density as:

$$P(x) = 2\pi^{-0.5m}\left|S_i^{-0.5}\right|\exp[-0.5(y^{'}S_i^{-1}y)]$$

where, $\left|S_i^{-0.5}\right|$ denotes the determinant of specified variance-covariance matrix for class i. The term $y^{'}S^{-1}y$ is the Mahalanobis distance used to estimate distance of each pixel from its class mean.

The function P(x) calculates the probability values so that the pixel having maximum probability can be allocated to its corresponding class. This expression can be simplified by taking logarithm to base $e$ given as:

$$\ln(P(x)) = -0.5p\ln(2\pi) - 0.5\ln|S| - 0.5(y^{'}S_i^{-1}y)$$

If both the sides be multiplied by -2 and if the constant terms of $p$ and $\ln(2\pi)$ are dropped, the expression becomes:

$$-\ln(P(x)) = \ln(|S|) + y^{'}S_i^{-1}y$$

The computation of probability now reduces to the derivation of Mahalanobis distance, the addition of logarithm of determinant of variance-covariance matrix and selection of minimum value from the results. The drawback of maximum likelihood classification is the large number of computations that must be carried out to classify each pixel especially when a large number of spectral channels are to be differentiated. The advantage of maximum likelihood classification is that a priori knowledge must be taken into account. By this we mean the a priori knowledge regarding the proportion of area to be classified which is covered by each class can be represented using a vector of prior probabilities.

**8. Accuracy of Classification**

Evaluation of classification results is an important process in the classification procedure. Traditionally, the accuracy is determined empirically by comparing with corresponding reference or ground data wherein the results are tabulated in the form of a square matrix known as confusion matrix. Ideal situation is represented by a diagonal matrix where only principal diagonal elements have non-zero values *i.e.* all areas of the image have been correctly classified . This is popularly known as the classification error matrix or confusion matrix or a contingency table. Error matrix summarizes how well the classification has been performed and how well it has categorized pixels corresponding to each land cover type. Within this error matrix, the known cover types used for training are represented along columns and the pixels actually classified into each land cover type are shown along rows. Several characteristics can be derived from error matrix. All non diagonal elements of the error matrix represent errors of omission or commission. Omission errors correspond to nondiagonal column elements while commission errors are represented by nondiagonal row elements. Overall classification accuracy is estimated by dividing the total number of correctly classified pixels by the total number of reference pixels. In a similar manner, the individual land use type classification accuracy can also be estimated.

Producer's accuracy is calculated by dividing the number of correctly classified pixels in each category by the number of training set pixels used for that category. User's accuracy is calculated by dividing the number of correctly classified pixels in each category by the total number of pixels that were classified in that category.

A multivariate technique for accuracy assessment derived using the error matrix is the Kappa statistic. Kappa statistic is a measure of agreement which can be computed using the expression:

$$K_{hat} = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} x_{i+} * x_{+i}}{N^2 - \sum_{i=1}^{r} x_{i+} * x_{+i}}$$

Where $r$ is the number of rows in the error matrix, $x_{ii}$ is the number of observations in row $i$ and column $i$, $x_{i+}, x_{+1}$ are the marginal totals for row $i$ and column $i$. And $N$ represents the total number of observations.