6. **SUPERVISED DATA ANALYSIS: A NEW DIRECTION**

"There is so much good in the worst of us; And so much bad in the best of us; That it ill becomes any of us To look down on the rest of us."

- C.W.L

### 6.1 Introduction

Multispectral remotely sensed imagery may be subject to either supervised or unsupervised analysis or, more typically, a hybrid of the two approaches. Unlike the unsupervised methods, supervised methods require some input from the user before the algorithm is initiated. The input is derived from field work, analysis of aerial photography or from the study of appropriate maps.

The basic strategy in supervised classification is to sample areas of known cover types to determine representative spectral values of each cover type. These sample areas are generally referred to as “training fields” and the representative spectral values from these training fields are sometimes called spectral signatures. Once representative spectral values have been established for each cover type, an image can then be classified. Each pixel is predicted or classified based on its similarity to the spectral values of representative cover type (Verbyla, 1995).

---

* Some parts of the material in this chapter appeared in the following research papers:
Supervised classification methods are based upon prior knowledge of the number and certain aspects of the statistical nature of the spectral classes (Mather, 1987). The statistical parameters extracted from the training set are used to formulate a classification rule (Gowda, 1984). The statistical characteristics of the classes that are to be estimated from the training sample pixels will depend on the method of supervised classification used. It is of crucial importance to ensure that the priori knowledge of the number and statistical characteristics of the classes is reliable.

The accuracy of a supervised classification analysis will depend upon two factors: i) the representativeness of the estimates of the number and statistical nature of the spectral classes present in the image data and ii) the degree of departure from the assumptions upon which the classification technique is based. These assumptions vary from one technique to another; in general, the most sophisticated techniques have the most stringent assumptions (Mather, 1987).

Supervised algorithms like Parallelepiped or box classifier, centroid method, maximum likelihood classifier, decision-tree classifier, and Minimum-distance classifier which is analogous to the k-means unsupervised technique etc., (Mather, 1987; Verbyla, 1995; Deer et al., 1996; Kohei Arai, 1992; Xiuping Jia and Richards, 1994) are available in literature. A brief overview of some of the aforesaid methods will be described in Section 6.2. All these algorithms require that the number of categories be specified in advance, and estimation of certain statistical characteristics of the classes to which the pixels are to be allocated are known. The proposed new symbolic supervised classification algorithm is presented in Section 6.4 and the experimental results are given in the section to follow.
6.2 Review of Existing Methodologies

6.2.1 Parallelepiped Classifier

The nonstatistical Parallelepiped classifier is most often used for qualitative scene analysis, particularly when a good interactive colour image display system is available for training the classifier (Mather, 1987; Verbyla, 1995). In a supervised mode of operation, the analyst locates a few regions in the image (e.g., forest fields, water bodies, cultivation land etc.,) representative of each of the apparent ground cover types. To define the feature-space Parallelepipeds corresponding to each ground cover class, the system determines the maximum and minimum data values in each of the specified spectral bands for each of the selected regions. More than one Parallelepiped may be required for a given class, depending on the degree of spectral heterogeneity of the class. Recognition proceeds by assigning a given pixel to the class having a Parallelepiped which contains that pixel based on proximity value.

This classifier has the advantages of being easy to train and able to classify a large area very rapidly. However, the results are highly prone to analyst's bias and it is possible, even likely, for pixels to be ambiguously classified into more than one class since Parallelepipeds of different classes may overlap. It is also possible for some pixels to lie outside all defined Parallelepipeds and thereby be impossible to classify.

6.2.2 Minimum-distance Classifier

The Minimum-distance classifier is the simplest of the statistical classifiers commonly used for remote sensing applications (Mather, 1987; Verbyla, 1995). Class means are determined from training data by using either supervised or unsupervised methods. Each pixel is then assigned to the class with the nearest
class mean. This classifier requires a moderate amount of training data, that is, sufficient to compute reliable estimates of multivariate class means, and the results obtained are not optimal if the class covariance matrices are not equal.

The major disadvantage with this classifier is that it uses only the spectral mean values instead of information about the spectral variability of the cover type. However, it is relatively fast compared to other statistical methods and does not suffer from the problems of the Parallelepiped classifier in allowing pixels to be classified ambiguously or not to be classified at all.

6.2.3 Maximum Likelihood Classifier

The Gaussian maximum-likelihood classifier is by far the classifier most commonly used in remote sensing (Mather, 1987; Verbyla, 1995; Kohei Arai, 1992; Xiuping Jia and Richards, 1994). The mean vectors and covariance matrices of the classes required to compute the class-conditional density functions as part of the recognition process are estimated by supervised or unsupervised methods. When it is determined that some classes grossly violate the underlying multivariate Gaussian (normal) assumption, the problem is usually dealt with by subdividing the offending classes into subclasses, each represented by its own mean vector and covariance matrix. In practice, the classes tend to have Gaussian distributions, and, moreover, the classifier is relatively tolerant to deviation from normality.

There are two principal disadvantages of this classifier. i) As second-order statistics are required, more training data are needed to characterize the classes adequately than are needed by the Minimum-distance classifier. ii) The recognition process is comparatively slow and the computation time varies approximately as the square of the number of features used (the variation is linear for the other two classifiers described). These can be some serious disadvantages because large areas often must be classified, but where operational use of this classifier is made, this problem is attacked through clever software implementation on an array processor.
Generalisations of the Gaussian maximum-likelihood classifier are employed. When reasonable estimates of the class prior probabilities are available, the classification may be based on maximising the posteriori probabilities, which are proportional to the product of the prior probabilities and the class-conditional density functions. When some classes in the scene are of greater importance than others, a minimum risk strategy may be adopted in the classification in favor of more important classes.

The drawbacks of the aforesaid traditional methods can be overcome in a better way by using a hybrid classifier which incorporates two or more decision rules (classification algorithms). In this paper, an efficient symbolic version of the hybrid classifier is proposed (Sect. 6.3.4) and the corresponding algorithmic details are given in Section 6.4.

6.3 Proposed Hybrid Methodology

6.3.1 Establishing Training Fields

It cannot be overemphasized that the success of applying Pattern Recognition to remote sensing data (as with any other type of data) depends critically on the use of an effective classifier training procedure. A training procedure is considered effective if it produces characterizations of the classes that are truly representatives of the scene to be classified. Often it is advisable that the training set should meet the general requirements like i) training classes should be mutually exclusive, ii) training area should be homogeneous, iii) training classes should be comprehensive, i.e., all the types are represented, iv) a sufficiently large number of pixels should be available for training (i.e., 10 to 100), and v) statistically, each class should exhibit a normal distribution (Gupta, 1991).

A training field is a simple area for estimating representative spectral statistics of a certain cover type. For example, if we are interested in forest type mapping
in the Coorg district of Karnataka State, India, we might establish training fields in the evergreen, semi-evergreen, deciduous, coniferous etc. Since the supervised algorithm rely on these training fields to estimate typical spectral values for each cover type, it is extremely important to establish training fields from homogeneous cover type areas.

Training samples are normally located by field work or from air photograph or map interpretation, and their positions on the image found either by visual inspection or by carrying out some geometric corrections (Mather, 1987) on the image to be classified. While training, generally a CRT device is used to display the scene and various classes are outlined on the screen as smaller windows to constitute training areas. Multispectral data pertaining to the delineated classes are screened and sorted out from data files by the computer for statistical analysis (Gupta, 1991). There are several approaches that can be used in establishing training fields including map digitization, on-screen digitizing, and a seed-pixel approach (Verbyla, 1995).

In map digitization training fields are commonly transferred from aerial photographs to base maps and then digitized from maps using a digitizing tablet. The potential problem with the map digitizing approach is that the training field cover types may differ between old maps or aerial photographs and recent satellite imagery. This disadvantage can be avoided by on-screen digitizing of training fields directly from the digital image. Training field polygon boundaries can be simply traced on the screen with most image processing systems. With the seed-pixel approach, one representative pixel is chosen as a starting pixel for training field delineation. Then candidate pixels around the seed pixel are sequentially considered as possible additional training-field pixels.

The proposed supervised classifier uses on-screen digitizing to establish initial training fields. In the later stages of the algorithm, method similar to the seed-pixel approach is used to improve the performance of the classifier.
6.3.2 Spectral Signatures

The training fields are used to estimate the variability of spectral values from each cover type. This spectral variability information can then be used in a classification rule. The supervised classifier proposed in this work requires the least information from the user. For each of the \( k \) classes specified (e.g., forest class, water class, plantation etc.), an estimation of the minimum and maximum pixel values on each \( d \)-dimensions from training field are utilized to represent the spectral signature of a specific land cover type. Alternatively a range, expressed in terms of a given number of standard deviation units on either side of the mean of each feature, can also be used. These values allow the estimation of the position of the boundaries of each spectral class. These range (minimum and maximum) values are treated as interval types of symbolic quantitative values. These quantitative interval type of values are used as spectral class signature of particular land cover type and they define regions of the \( d \)-dimensional feature space identified with particular land cover type (or information class).

6.3.3 Formulation of Composite Symbolic Objects

Merging is the process of gathering together, based on a similarity measure of two samples and assigning them same cluster membership, or label for further clustering. In symbolic data analysis, the concept of Composite Symbolic Object (CSO) is used to describe a newly formed object resulted when two symbolic objects are merged (for more detail, refer Sect. 4.3.4).

The spectral response patterns measured by remote sensors are variable in nature over different geographic area for a particular class. Hence, it is wise to consider the spectral variation of candidate pixels around the training-field as possible additional spectral class signature values. This modification in training class boundaries allows to grab the pixels that are almost close to the expected cover type. The major advantage of updating the spectral class signature values is that
it reduces the classification error rate through strengthening the training field boundary. For updating the spectral class signature values, a new CSO is formed at every stage of the assigning process of a pixel to its nearest spectral class by considering the pixel feature values and its spectral class signature values. This new spectral class signature value (CSO) is used for further classification. This new method of formulation of CSO is described quantitatively in Section 4.3.4.

### 6.3.4 Symbolic Supervised Classifier

In this section, an efficient symbolic version of the hybrid classifier is proposed. It utilizes the Parallelepiped rule to select representative spectral class signatures of well-separated classes and a rule similar to the Minimum-distance classifier to assign the pixels to one of the nearest spectral class signature.

A supervised symbolic hybrid classifier was developed to operate over arbitrary image sizes having multiple bands and different ground cover categories. Training fields are selected through the aid of on-screen digitization by a human operator. For establishing the spectral class signature (training class boundary), the classifier utilizes the property of Parallelepiped learning. Unlike the spectral class mean values computed for the selected regions, the proposed classifier uses the information about spectral variability of each cover class (lowest and highest pixel values in each band or feature). It characterizes the spectral boundary values for each class as symbolic quantitative type of interval values.

After establishing the spectral values for each cover type, a methodology which is similar to the Minimum-distance rule is employed to classify the image data set. The symbolic distance from each unknown pixel is calculated for each spectral class signature and the pixel is predicted or classified by assigning it to the symbolic spectral class signature (representative cover type spectral values) having the highest similarity. The symbolic similarity measure is used here to associate image pixels to the nearest class signature more accurately (see Sect. 4.3.2).
Although, the information contained in the training field is just sufficient to classify the pixels present in the data set, it is wise to update the information content of training sets, whenever a new pixel is assigned to it. The proposed method of formulation of CSO is used in this regard to agglomerate the pixel feature values with its nearest spectral class signature values (Sect. 6.3.3). Because of the aforesaid rules, the performance of the algorithm is more effective when compared to the other classifiers.

6.4 Computational Algorithm

Let the remotely sensed data consist of 'm' number of pixels and 'd' number of dimensions. These data, or the actual measurements, are seldom used just as they are recorded unless a probabilistic model for pattern generation is available. Some normalization is usually employed based on the requirements of the analysis (Jain and Dubes, 1988). Preparing the data for a cluster analysis requires some sort of normalization that takes into account the measure of proximity. Hence, as a preliminary step (Pass # 1), the remotely sensed data are normalized and interpreted as symbolic quantitative interval type of feature values.

The block diagram given in Fig. 6.1 illustrates the major stages involved in supervised data analysis. The algorithm, comprising of three passes, is given below.

**Pass # 1:** Remotely sensed data is pre-processed during this pass.

1. Determine minimum and maximum feature values in each dimension.
2. Using the minimum and maximum values determined in step 1, normalize the feature vectors between 1 and $n$ (user defined limit).
3. Determine symbolic quantitative feature values from the normalized feature vectors.
Figure 6.1 Block diagram illustrating Supervised data analysis
Pass # 2: Spectral class signatures are identified during this pass.

4. Establish the training fields through on-screen digitization and seed-pixel approaches for the \( k \) specific cover types.

5. Determine the \( k \) number of symbolic spectral class signature (see Sect. 4.2) from the established training information.

Pass # 3: The initial \( m \) number of pixels are classified into \( k \) number of classes during this pass.

6. Begin with \( k \) number of symbolic spectral class signature and \( m \) number of pixels.

7. For \( i = 1 \) to \( m \), perform steps 8 to 11.

8. For \( j = 1 \) to \( k \), find the similarity distance measures \( S \) between \( i \)th sample and \( j \)th symbolic spectral class signature.

9. Classify the pixel \( i \) by assigning it to the symbolic spectral class signature having the highest similarity.

10. Determine the composite symbolic object by considering pixel \( i \) and its spectral class signature.

11. Store the classification information.

6.5 Experimentation

Experiments are carried out to authenticate the superiority of the proposed hybrid classifier and the results are given below. The computer simulated data sets as well as real multispectral remotely sensed data are used for the experimental purpose. As the proposed symbolic hybrid classifier is based on two well-known methods viz., the Parallelepiped classifier and the Minimum-distance classifier, the results obtained from the proposed method are compared and contrasted with results from the Parallelepiped classifier alone and the Minimum-distance classifier.
alone. The display scheme illustrated in Appendix A.3 is employed in this chapter to plot the output classification maps.

**Experiment No. 1:**

In this experiment, the computer simulated randomly generated data is used to substantiate the efficacy of the proposed classifier and to illustrate some drawbacks presented in the conventional (Parallelepiped and Minimum-distance) classifiers. The simulated data used in this experiment is shown in the Fig. A.4.1 (given in Appendix A.4). This simulated image is of size 512 X 512 with three features. The number of classes generated and their mean feature values are given in Table 6.1. Figure 6.2a shows the classification map obtained from the proposed hybrid classifier. Figure 6.2b and Fig. 6.2c show the classification map from the Parallelepiped classifier and the Minimum-distance classifier respectively. Table 6.2 shows the classification results obtained from the proposed symbolic hybrid, Parallelepiped, and Minimum-distance classifiers.

The results show that, there is some misclassification in Parallelepiped and Minimum-distance classifiers. Due to the overlap of spectral class signature values in the Parallelepiped classifier, some samples (pixels with black colour depicted in Fig. 6.2b) are ambiguously classified into more than one class and some samples are left unlabeled as they lie outside the defined Parallelepipeds. In the Minimum-distance classifier, as it uses conventional distance measures and only spectral mean values (instead of information about the spectral variability of the cover type), some samples have been incorrectly labelled due to equidistance of samples to two or more spectral classes. Where as, the proposed classifier overcomes these drawbacks and it is ascertained from the results which perfectly match with the generated data set.
Experiment No. 2:

The proposed symbolic hybrid classifier is applied on IRS 1A LISS-II B1 satellite data covering forest area of Coorg District, Karnataka State, India. The study area is sensed in the path/row 27-60. The data acquired in four different spectral ranges has a resolution of 36.25 mt. The registered IRS satellite data shown in Fig. A.4.4 (given in Appendix A.4) is acquired on 03 Feb. 1991 and is used in this experiment. The standard scene is registered based on the reference point and it covers 1600 X 1300 samples of the standard scene (size 2520 X 2500). The chosen training sets represent each of the seven classes 'evergreen forest', 'deciduous forest', 'coffee plantation', 'paddy', 'water bodies', 'crop land', and 'rubber' to classify the entire study area.

The corresponding classification map from the proposed method is shown in Fig. 6.3a. Figure 6.3b and Fig. 6.3c shows the classification map of the Parallelepiped classifier and the Minimum-distance classifier respectively. Due to the overlap of spectral class signature values in the Parallelepiped classifier, some samples (pixels with black colour depicted in Fig. 6.3b) are ambiguously classified into more than one class and some samples are left unlabeled as they lie outside the defined Parallelepipeds. Piechart given in the Fig. 6.4a through Fig. 6.4c shows the percentage of area covered by different land covers. The class legends, their cover types and other related information are presented in Table 6.3.

The results obtained from this experiment has been verified with the aid of reference maps, the ground truth information collected from the respective forest divisions, irrigation departments and other field experts. Topographic maps prepared by the Survey of India, and Institute of Pondicherry are used as reference maps in this work. The land use/land cover map prepared by Karnataka State Remote Sensing Technology Utilization Centre, Dept. of Science and Technology, Bangalore, India is also used in assigning the ground cover type to the main clusters.
6.6 Critical and Comparative Analysis

It is instructive to compare the proposed methods with some other methods to justify the superiority of the proposed hybrid classifier. As a preliminary step, the proposed classifier is judiciously compared with the well-known Parallelepiped and Minimum-distance classifiers in terms of memory and time requirements. In the following paragraphs, some comparative studies made with other schemes are presented.

The Parallelepiped and Minimum-distance classifiers are simple and fast as compared with the Maximum Likelihood classifier. The Maximum Likelihood technique uses more computational time due to the computation of mean vectors and covariance matrices. The proposed hybrid classifier almost takes the same amount of computation time as required by the Parallelepiped and Minimum-distance classifier as these algorithms are based on pixel-by-pixel classification. Due to the updation of training field boundaries in the proposed classifier, additional computational time required for arithmetic operations is negligible.

As far as the performance of the classifier is concerned, the proposed hybrid classifier is better than the conventional Parallelepiped and Minimum-distance classifier and is ascertained from the experimental results presented in the previous section. Some disadvantages of Parallelepiped and Minimum-distance classifiers are highlighted in the Section 6.2.

The estimation of classification accuracy was subjected to a two-pronged test. The amount of confusion with different training sets defined in the groundtruth and the classes in the classified sub-scene was estimated by generating a confusion matrix to check the overall classification accuracy (Sect. 8.4.1). Finally, a statistical analysis technique (Kappa coefficient) is computed from the confusion matrix (Sect. 8.4.2) and is utilized to demonstrate the classification accuracy. The larger its value, the more accurate the classification results. The verification results are highlighted in chapter 8 (Section 8.5).
6.7 Summary

Symbolic hybrid classifier incorporating two or more decision rules is proposed for classifying remotely sensed multispectral data. The proposed symbolic supervised classifier is a hybrid of two well-known methods called 'Parallelepiped' and 'Minimum-distance' classifiers. Training fields are established through on-screen digitization and seed-pixel approaches. The proposed new non-metric symbolic similarity measure is used to assign the pixels to the nearest class signature more accurately. A novel method of formulation of CSO proposed in the previous chapter is employed to strengthen the training field boundaries. Experiments are conducted using the IRS (Indian Remote Sensing) satellite data to demonstrate the capabilities of the proposed hybrid algorithm. To justify the superiority of the hybrid classifier, the results are compared against the results obtained from the Parallelepiped classifier alone and the results obtained from the Minimum-distance classifier alone. The overall classification accuracy is estimated by generating a Confusion matrix and computing Kappa index.
Figure 6.2a Classification results of randomly generated image
- Proposed symbolic hybrid classifier

Figure 6.2b Classification results of randomly generated image
- Conventional Parallelepiped classifier
Figure 6.2c  Classification results of randomly generated image
- Conventional Minimum-distance classifier
Supervised Data Analysis: A New Direction

Figure 6.3a Classification map of Coorg forest area, year 03 Feb. 1991
- Proposed symbolic hybrid classifier

Figure 6.4a Piehart showing the percentage of land covers of Coorg forest area, year 03 Feb. 1991 (Proposed symbolic hybrid classifier)
Supervised Data Analysis: A New Direction

Figure 6.3b Classification map of Coorg forest area, year 03 Feb. 1991
- Conventional Parallelepiped classifier

Figure 6.4b Piechart showing the percentage of land covers of Coorg forest area, year 03 Feb. 1991 (Parallelepiped classifier)
Supervised Data Analysis: A New Direction

Figure 6.3c Classification map of Coorg forest area, year 03 Feb. 1991
- Conventional Minimum-distance classifier

Figure 6.4c Piechart showing the percentage of land covers of Coorg forest area, year 03 Feb. 1991 (Minimum-distance classifier)
Table 6.1 Randomly generated image details

<table>
<thead>
<tr>
<th>Class No.s</th>
<th>No. of samples generated</th>
<th>Mean feature values of generating samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>108577</td>
<td>220, 150, 15</td>
</tr>
<tr>
<td>2</td>
<td>27102</td>
<td>10, 50, 20</td>
</tr>
<tr>
<td>3</td>
<td>17858</td>
<td>20, 215, 70</td>
</tr>
<tr>
<td>4</td>
<td>16584</td>
<td>50, 220, 50</td>
</tr>
<tr>
<td>5</td>
<td>33963</td>
<td>150, 115, 50</td>
</tr>
<tr>
<td>6</td>
<td>58060</td>
<td>250, 40, 15</td>
</tr>
</tbody>
</table>

Table 6.2 Classification results

<table>
<thead>
<tr>
<th>Number of samples predicted by the classification algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class No.s</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>Mis-classification</td>
</tr>
</tbody>
</table>
Table 6.3 Classification results

<table>
<thead>
<tr>
<th>Class No.s</th>
<th>Legends</th>
<th>Type of ground covered</th>
<th>Spectral signature values of four bands (b1 to b4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Crop Land</td>
<td>32-40; 28-33; 25-28; 38-42</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Water</td>
<td>10-15; 11-17; 16-18; 30-33</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Paddy</td>
<td>37-44; 26-34; 22-32; 38-47</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Coffee</td>
<td>35-45; 12-15; 16-18; 30-32</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Deciduous Forest</td>
<td>32-36; 14-18; 18-20; 32-34</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Evergreen Forest</td>
<td>33-53; 10-13; 16-18; 28-30</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Rubber</td>
<td>18-23; 14-17; 16-18; 30-32</td>
</tr>
</tbody>
</table>

Note: colour ▲ represents mis-classification in the classification map.

Table 6.3 Continued

<table>
<thead>
<tr>
<th>Class No.s</th>
<th>Proposed hybrid classifier</th>
<th>Conventional Parallelepiped classifier</th>
<th>Conventional Minimum-distance classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of pixels</td>
<td>Area in Km²</td>
<td>No. of pixels</td>
</tr>
<tr>
<td>1</td>
<td>118980</td>
<td>156.35</td>
<td>11845</td>
</tr>
<tr>
<td>2</td>
<td>8825</td>
<td>11.60</td>
<td>6768</td>
</tr>
<tr>
<td>3</td>
<td>148281</td>
<td>194.85</td>
<td>73344</td>
</tr>
<tr>
<td>4</td>
<td>734114</td>
<td>964.67</td>
<td>339073</td>
</tr>
<tr>
<td>5</td>
<td>501602</td>
<td>659.14</td>
<td>93897</td>
</tr>
<tr>
<td>6</td>
<td>535316</td>
<td>703.44</td>
<td>328617</td>
</tr>
<tr>
<td>7</td>
<td>32882</td>
<td>43.21</td>
<td>996</td>
</tr>
</tbody>
</table>

Mis-classification 1225460