Chapter 11 Image Processing - Classification

11.1 Classification Techniques

Classification of remotely sensed data is used to assign corresponding levels with respect to groups with homogeneous characteristics, with the aim of discriminating multiple objects from each other within the image.

The level is called class. Classification will be executed on the base of spectral or spectrally defined features, such as density, texture etc. in the feature space. It can be said that classification divides the feature space into several classes based on a decision rule. Figure 11.1.1 shows the concept of classification of remotely sensed data.

In many cases, classification will be undertaken using a computer, with the use of mathematical classification techniques. Classification will be made according to the following procedures as shown in Figure 11.1.2.

Step 1: Definition of Classification Classes
Depending on the objective and the characteristics of the image data, the classification classes should be clearly defined.

Step 2: Selection of Features
Features to discriminate between the classes should be established using multi-spectral and/or multi-temporal characteristics, textures etc.

Step 3: Sampling of Training Data
Training data should be sampled in order to determine appropriate decision rules. Classification techniques such as supervised or unsupervised learning will then be selected on the basis of the training data sets.

Step 4: Estimation of Universal Statistics
Various classification techniques will be compared with the training data, so that an appropriate decision rule is selected for subsequent classification.

Step 5: Classification
Depending up on the decision rule, all pixels are classified in a single class. There are two methods of pixel by pixel classification and per-field classification, with respect to segmented areas.

Popular techniques are as follows.
a. Multi-level slice classifier
b. Minimum distance classifier
c. Maximum likelihood classifier
d. Other classifiers such as fuzzy set theory and expert systems

Step 6: Verification of Results
The classified results should be checked and verified for their accuracy and reliability.

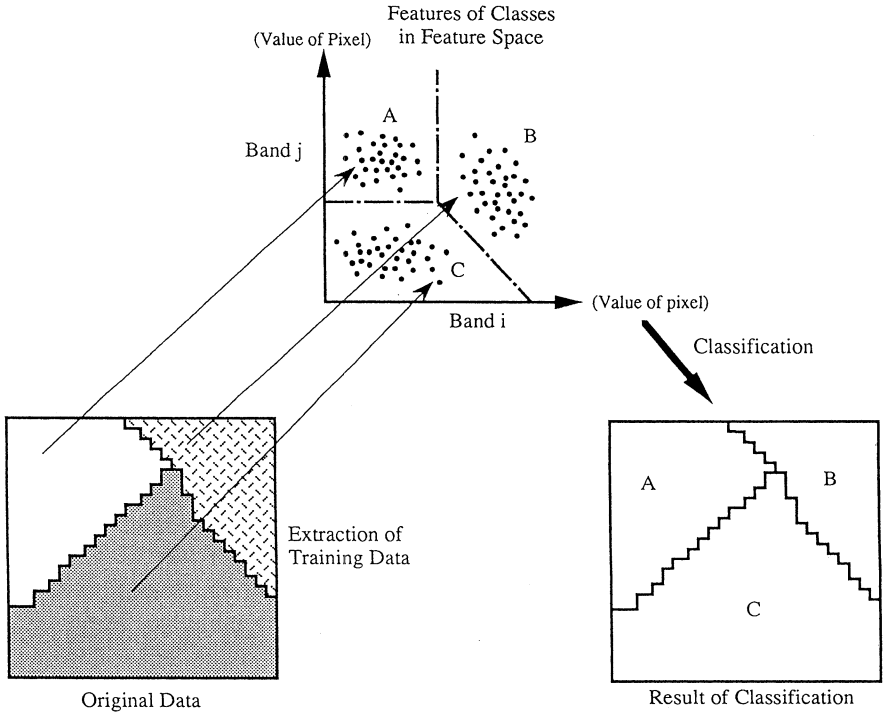


Figure 11.1.1 Concept of Classification of Remotely Sensed Data

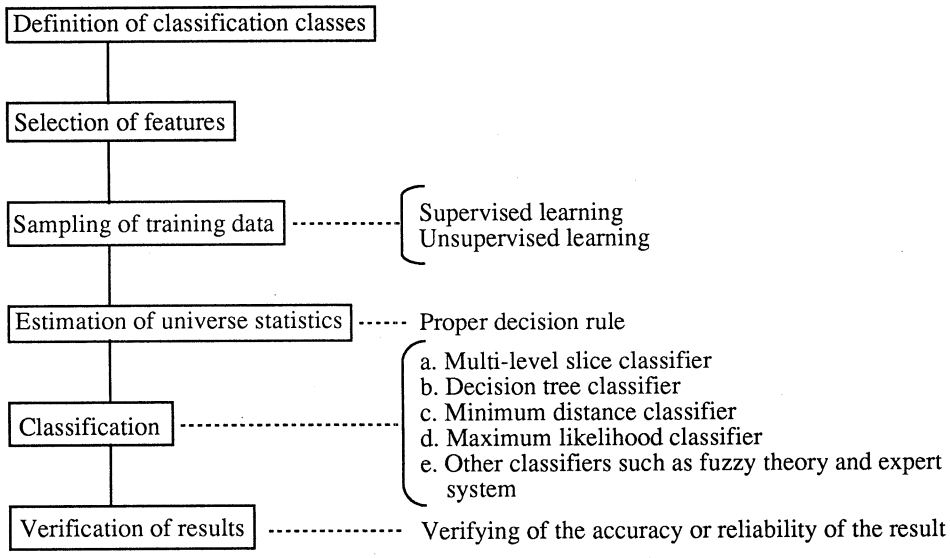


Figure 11.1.2 Procedures of Classification

11.2 Estimation of Population Statistics

a. **Supervised classification**

In order to determine a decision rule for classification, it is necessary to know the spectral characteristics or features with respect to the population of each class. The spectral features can be measured using ground-based spectrometers. However due to atmospheric effects, direct use of spectral features measured on the ground are not always available. For this reason, sampling of **training data** from clearly identified training areas, corresponding to defined classes is usually made for estimating the population statistics (see Figure 11.2.1). This is called supervised classification. Statistically unbiased sampling of training data should be made in order to represent the population correctly.

b. **Unsupervised Classification**

In the case where there is less information in an area to be classified, only the image characteristics are used as follows.

(1) Multiple groups, from randomly sampled data, will be mechanically divided into homogeneous spectral classes using a clustering technique (see 11.3).
(2) The clustered classes are then used for estimating the population statistics. This classification technique is called unsupervised classification (see Figure 12.2.2).

c. Estimation of Population Statistics

Maximum likelihood estimation is the most popular method by which the population statistics such as mean and variance, are estimated to maximize the probability or likelihood from a defined probability density function within the feature space.

In most cases, the probability density function is selected to be a multiple normal distribution. The multiple normal distribution gives the following the maximum likelihood estimator.

$$\text{Mean} ; \mu ei = \frac{1}{n}\sum_{j=1}^{n} X_{ij} \quad (i = 1,2, \ldots m)$$

Variance - covariance matrix

$$\Sigma_e = \frac{1}{m}\sum_{i=1}^{m} (X_i - \mu_e)(X_i - \mu_e)$$

where m: number of bands
n: number of pixels

Before adopting the maximum likelihood classification, it should be checked to determine if the distribution of training data will fit the normal distribution or not. (see Figure 12.2.3)

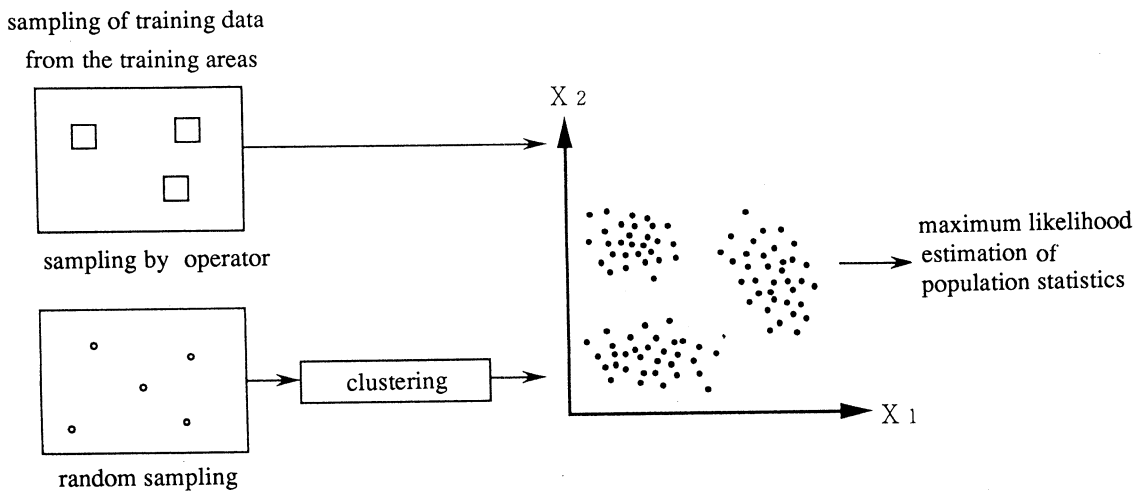


Figure 11.2.1 Sampling of Training Data by Operator and Clustering

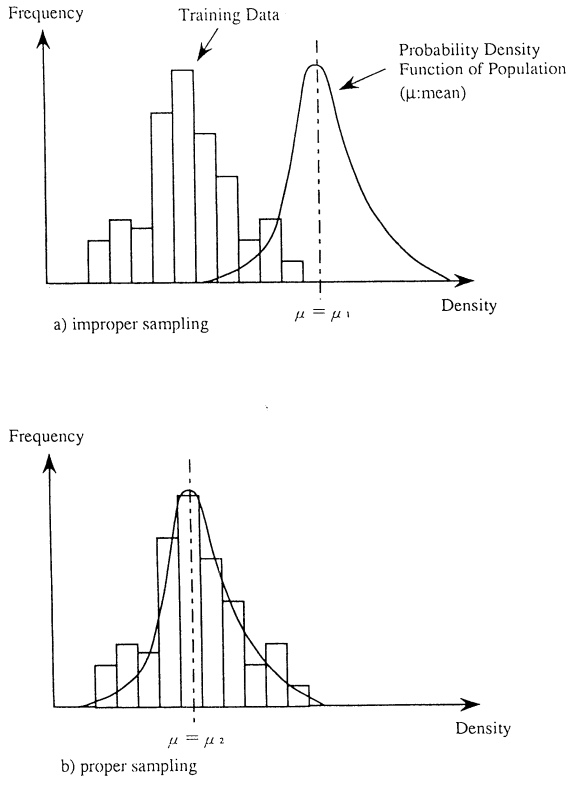


Figure 11.2.2 Maximum Likelihood Estimation of Population Statistics

11.3 Clustering

Clustering is a grouping of data with similar characteristics. Clustering is divided into hierarchical clustering and non-hierarchical clustering as mentioned as follows.

a. Hierarchical Clustering

The similarity of a cluster is evaluated using a "distance" measure. The minimum distance between clusters will give a merged cluster after repeated procedures from a starting point of pixel-wise clusters to a final limited number of clusters.

Figure 11.3.1 shows the general procedure of hierarchical clustering.

The distances to evaluate the similarity are selected from the following methods.

(1) Nearest neighbor method
Nearest neighbor with minimum distance will form a new merged cluster.
(2) Furthest neighbor method
Furthest neighbor with maximum distance will form a new merged cluster.
(3) Centroid method
Distance between the gravity centers of two clusters is evaluated for merging a new merged cluster.
(4) Group average method
Root mean square distance between all pairs of data within two different clusters, is used for clustering.
(5) Ward method
Root mean square distance between the gravity center and each member is minimized.

b. Non-hierarchical Clustering

At the initial stage, an arbitrary number of clusters should be temporally chosen. The members belonging to each cluster will be checked by selected parameters or distance and relocated into the more appropriate clusters with higher separability. The ISODATA method and K-mean method are examples of non-hierarchical clustering.

The ISODATA method is composed of the following procedures.
(1) All members are relocated into the closest clusters by computing the distance between the member and the clusters.
(2) The center of gravity of all clusters is recalculated and the above procedure is repeated until convergence.
(3) If the number of clusters is within a certain specified number, and the distances between the clusters meet a prescribed threshold, the clustering is considered complete.

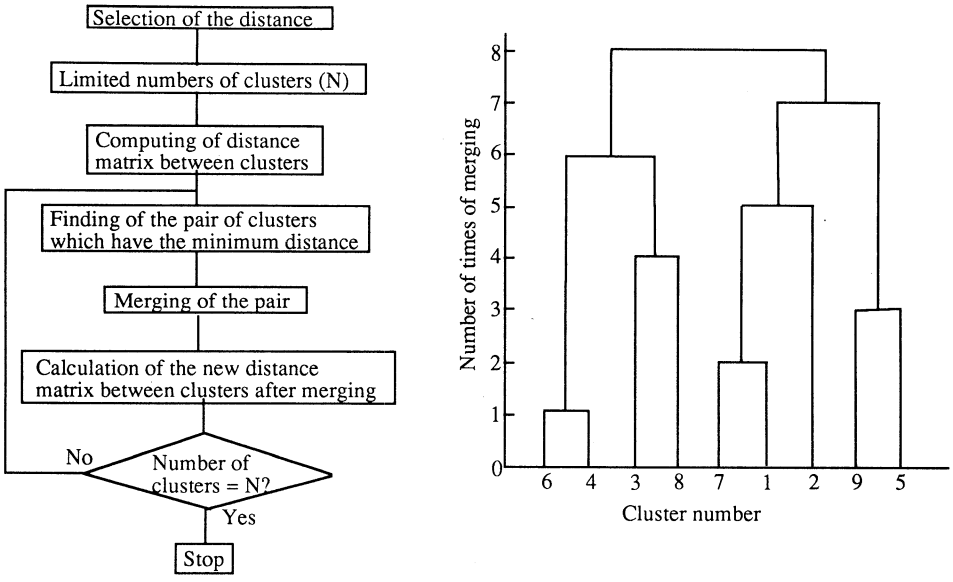


Figure 11.3.1 Flow and an example of hierarchical clustering

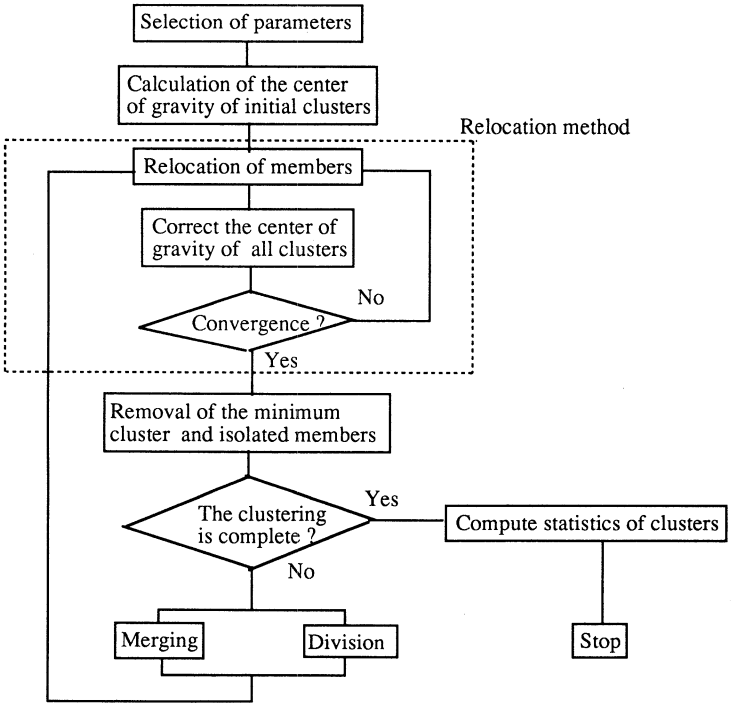


Figure 11.3.2 Flow of non-hierarchical clustering (ISODATA method)

11.4 Parallelpiped Classifier

The **parallelpiped classifier** (often termed **multi-level slicing**) divides each axis of multi-spectral feature space, as shown in an example in Figure 11.4.1. The decision region for each class is defined on the basis of a lowest and highest value on each axis. The accuracy of classification depends on the selection of the lowest and highest values in consideration of the population statistics of each class. In this respect, it is most important that the distribution of population of each class is well understood.

The parallelpiped classifier is very simple and easy to understand schematically. In addition the computing time will be a minimum, when compared with other classifiers.

However the accuracy will be low especially when the distribution in feature space has covariance or dependency with oblique axes. Orthogonalization should be undertaken using principal component analysis, for example, before adopting the parallelpiped classifier.

Figure 11.4.2 shows an example of classification with the use of the parallelpiped classifier.

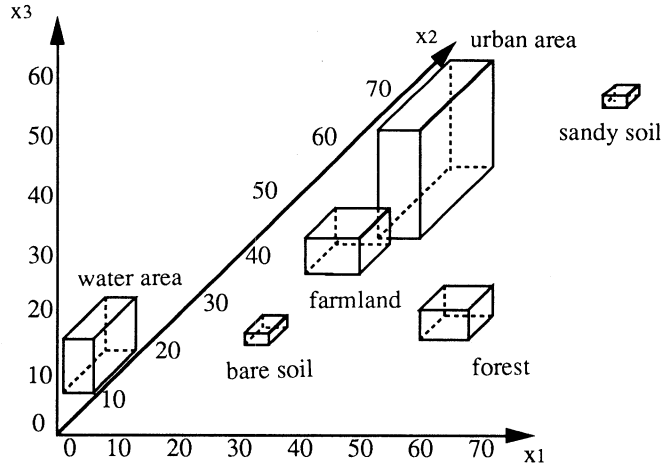


Figure 11.4.1 Schematic Concept of Parallel Piped Classifier in Three Dimensional Feature Space

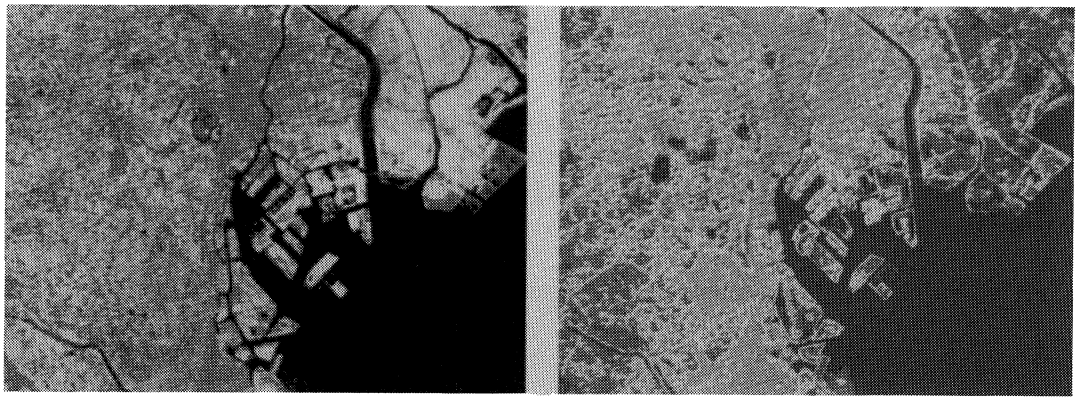


Figure 11.4.2 Example of Classification with use of Parallel Piped Classifier

11.5 Decision Tree Classifier

The **decision tree classifier** is an hierarchically based classifier which compares the data with a range of properly selected features. The selection of features is determined from an assessment of the spectral distributions or separability of the classes. There is no generally established procedure. Therefore each decision tree or set of rules should be designed by an expert. When a decision tree provides only two outcomes at each stage, the classifier is called a binary decision tree classifier (BDT).

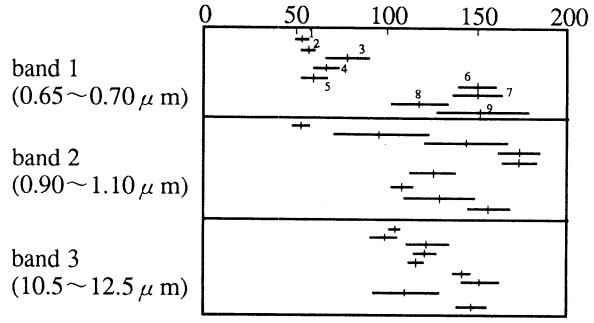
Figure 11.5.1 shows the spectral characteristics of ground truth data for nine classes and the corresponding decision tree classifier to classify the nine classes using their spectral characteristics.

Generally a group of classes will be classified into two groups with the highest separability with respect to a feature.

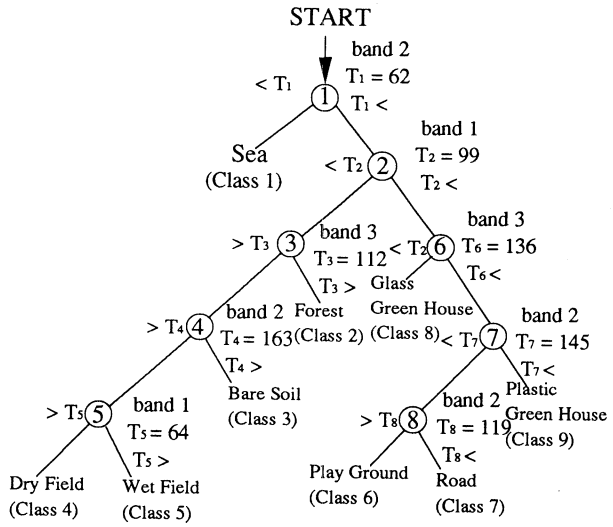
Features often used are as follows.
(1) Spectral values
(2) An index which is computed from spectral values. For example, the vegetation index is a popular indices.
(3) any arithmetic value such as addition, subtraction or ratioing.
(4) Principal components.

The advantages of the decision tree classifier are that computing time is less than the maximum likelihood classifier and by comparison the statistical errors are avoided. However the disadvantage is that the accuracy depends fully on the design of the decision tree and the selected features.

Figure 11.5.2 shows an example of classification with a decision tree classifier.



a) Spectral characteristics of nine classes



b) Decision Trees

Figure 11.5.1 Hierarchical Classification by Decision Tree Classifier

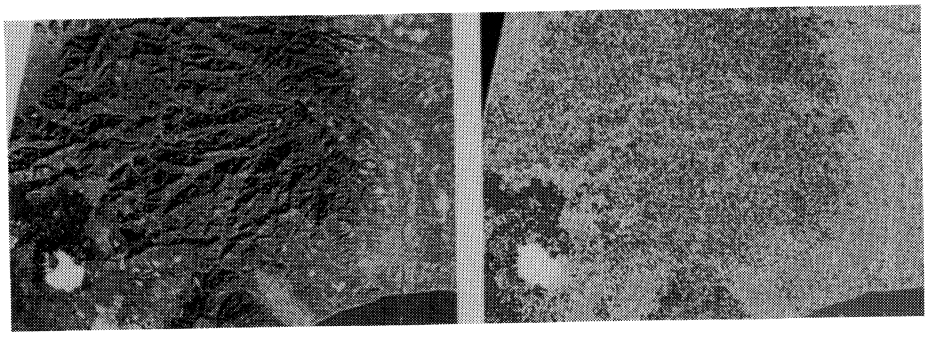


Figure 11.5.2 Example of Classification withDecision Tree classifier

11.6 Minimum Distance Classifier

The **minimum distance classifier** is used to classify unknown image data to classes which minimize the distance between the image data and the class in multi-feature space. The distance is defined as an index of similarity so that the minimum distance is identical to the maximum similarity. Figure 11.6.1 shows the concept of a minimum distance classifier. The following distances are often used in this procedure.

(1) **Euclidian distance**

$$d_k^2 = (\mathbf{X} - \mu_k)^t . (\mathbf{X} - \mu_k)$$

Is used in cases where the variances of the population classes are different to each other. The Euclidian distance is theoretically identical to the similarity index.

(2) **Normalized Euclidian distance**

The Normalized Euclidian distance is proportional to the similarity in dex, as shown in Figure 11.6.2, in the case of difference variance.

$$d_k^2 = (\mathbf{X} - \mu_k)^t . \sigma_k^{-1} . (\mathbf{X} - \mu_k)$$

(3) **Mahalanobis distance**

In cases where there is correlation between the axes in feature space, the Mahalanobis distance with variance-covariance matrix, should be used as shown in Figure 11.6.3.

$$d_k^2 = (\mathbf{X} - \mu_k)^t . \Sigma_k^{-1} . (\mathbf{X} - \mu_k)$$

where

$\mathbf{X}$: vector of image data (n bands) $\quad \mathbf{X} = [x_1, x_2, \ldots x_n]$

μ_k : mean of the kth class $\quad \mu_k = [m_1, m_2, \ldots m_n]$

σ_k : variance matrix

$$\sigma_k = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & \cdots & \sigma_{nn} \end{bmatrix}$$

Σ_k : variance-covariance matrix

$$\Sigma_k = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

Figure 11.6.4 shows examples of classification with the three distances.

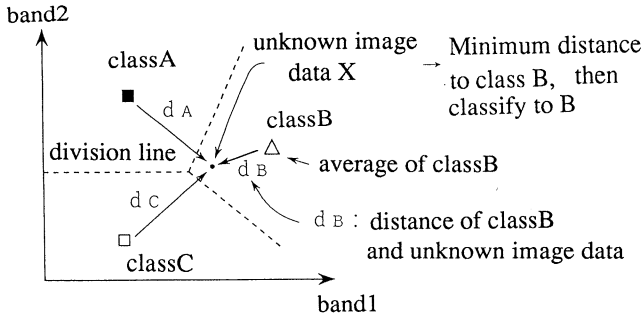


Figure 11.6.1 Concept of Minimum Distance Classifier

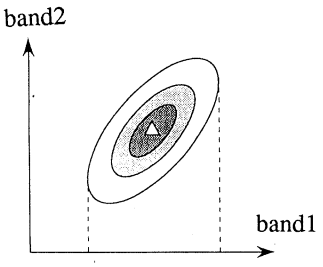
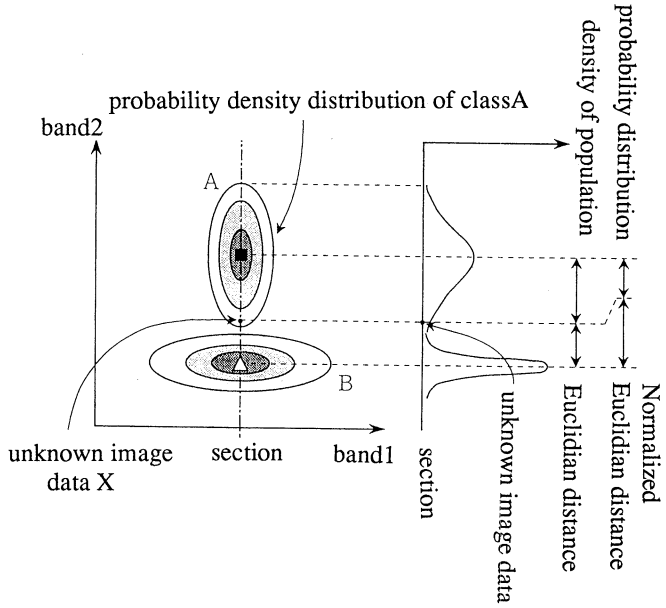


Figure 11.6.3 In Case When There is Correlation Amang Axes in Feature Space



Unknown pixel data x is nearer to class B by Euclidian distance,but it's better to classify to class A when Normalized Euclidian distance is used.

Figure 11.6.2 Normalized Euclidian distance

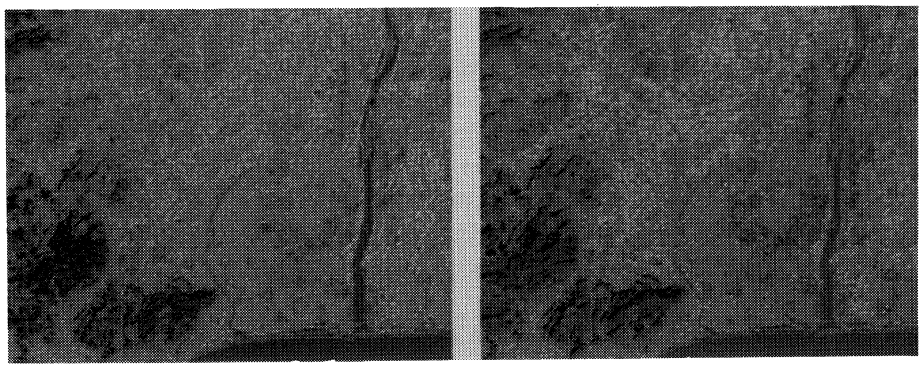


Figure 11.6.4 Example of Classification with Minimum Distance Classifier

11.7 Maximum Likelihood Classifier

The **maximum likelihood classifier** is one of the most popular methods of classification in remote sensing, in which a pixel with the maximum likelihood is classified into the corresponding class. The **likelihood** Lk is defined as the posterior probability of a pixel belonging to class k.

$$L_k = P(k/\mathbf{X}) = P(k)*P(\mathbf{X}/k) / \Sigma P(i)*P(\mathbf{X}/i)$$

where P(k) : prior probability of class k
P(X/k) : conditional probability to observe **X** from class k, or probability density function

Usually P(k) are assumed to be equal to each other and $\Sigma P(i)*P(\mathbf{X}/i)$ is also common to all classes. Therefore L_k depends on $P(\mathbf{X}/k)$ or the probability density function.

For mathematical reasons, a multivariate normal distribution is applied as the probability density function. In the case of normal distributions, the likelihood can be expressed as follows.

$$L_k(\mathbf{X}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\mu_k)\Sigma_k^{-1}(\mathbf{X}-\mu_k)^t\right\}$$

where n: number of bands
X : image data of n bands
$L_k(\mathbf{X})$: likelihood of **X** belonging to class k
μ_k : mean vector of class k
Σ_k : **variance-covariance matrix** of class k
$|\Sigma_k|$: determinant of Σ_k

In the case where the variance-covariance matrix is symmetric, the likelihood is the same as the Euclidian distance, while in case where the determinants are equal each other, the likelihood becomes the same as the Mahalanobis distances. Figure 11.7.1 shows the concept of the maximum likelihood method.

The maximum likelihood method has an advantage from the view point of probability theory, but care must be taken with respect to the following items.

(1) Sufficient ground truth data should be sampled to allow estimation of the mean vector and the variance-covariance matrix of population.
(2) The inverse matrix of the variance-covariance matrix becomes unstable in the case where there exists very high correlation between two bands or the ground truth data are very homogeneous. In such cases, the number of bands should be reduced by a principal component analysis.
(3) When the distribution of the population does not follow the normal distribution, the maximum likelihood method cannot be applied.

Figure 11.7.2 shows an example of classification by the maximum likelihood method.

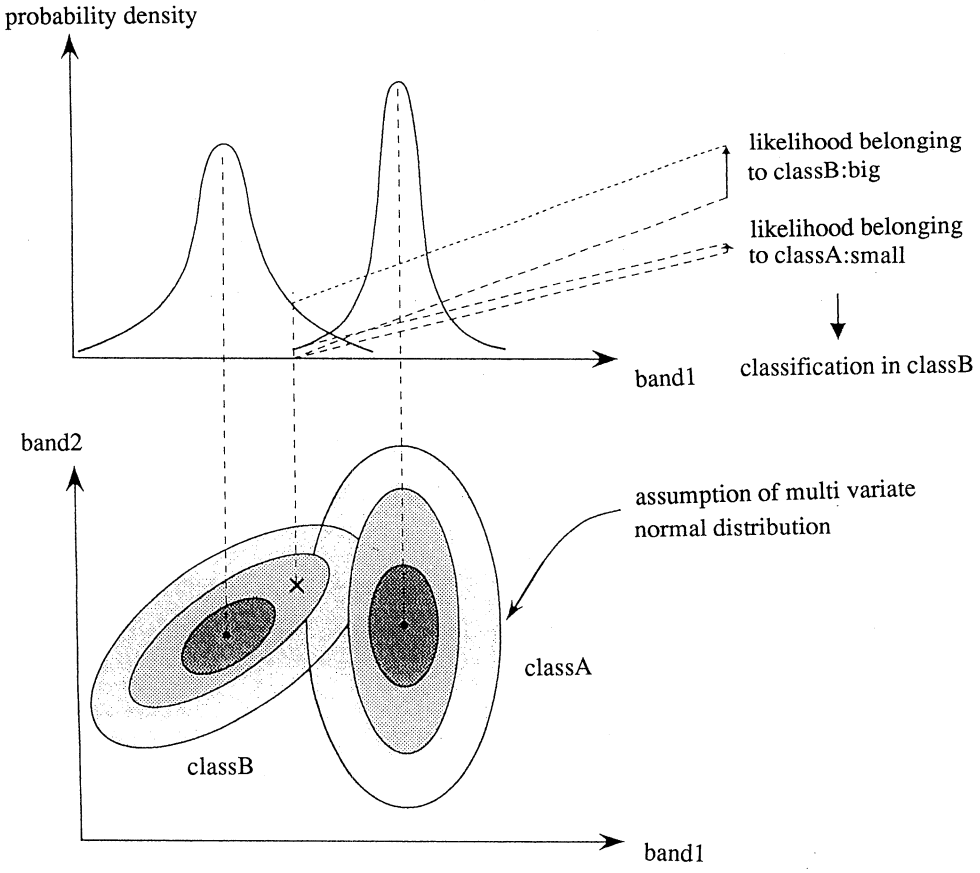


Figure 11.7.1 Concept of Maximum Likelihood Method

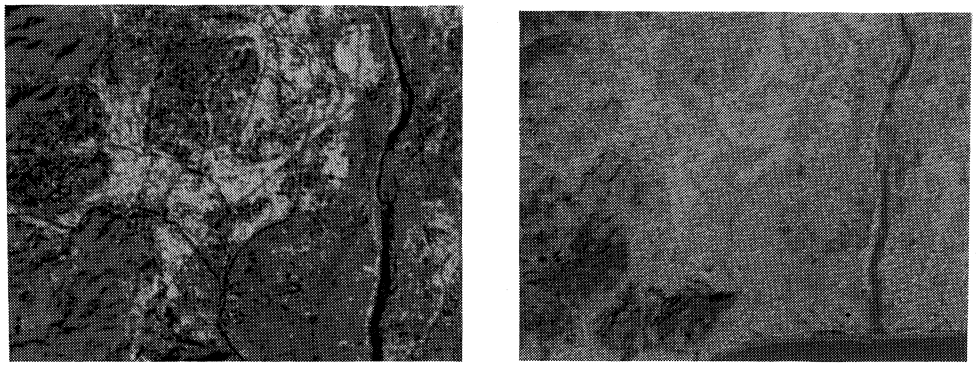


Figure 11.7.2 Example of Classification with Maximum Likelihood Method

11.8 Applications of Fuzzy Set Theory

Fuzzy set theory, to treat fuzziness in data, was proposed by Zadeh in 1965. In Fuzzy set theory the membership grade can be taken as a value intermediate between 0 and 1 although in the normal case of set theory membership the grade can be taken only as 0 or 1.

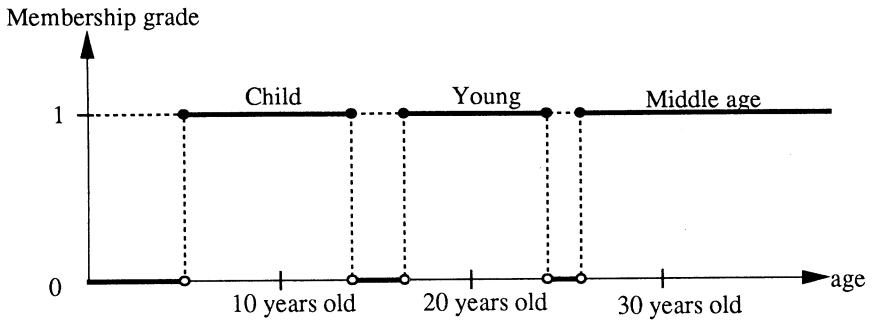
Figure 11.8.1 shows a comparison between the normal case of set theory and fuzzy set theory. The function of the membership grade is called its "membership function" in Fuzzy theory. The membership function will be defined by the user in consideration of the fuzziness.

In remote sensing it is often not easy to delineate the boundary between two different classes. For example, there are transitive vegetation or mixed vegetation between forest and grass land. In such cases as unclearly defined class boundaries, Fuzzy set theory can be usefully applied, in a qualitative sense.

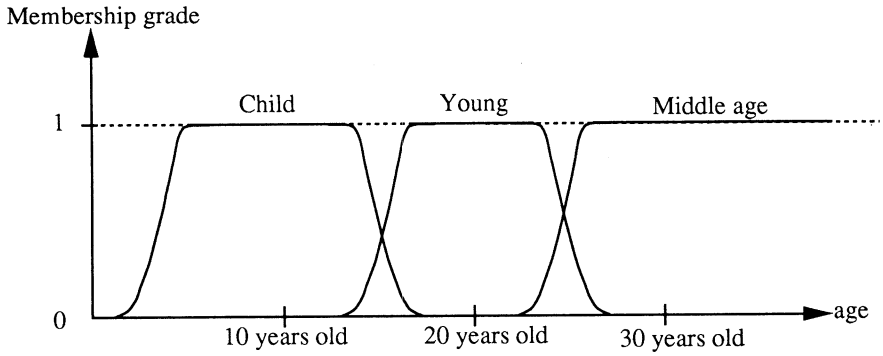
The following shows how the maximum likelihood method with Fuzzy set theory. Let the membership function be $Mf(\kappa)$ of class k (k=1,n), the likelihood Lf of fuzzy class f can be defined as follows.

$$L_f=\sum_{k=1}^{n}\left\{Mf(\kappa)*P(\mathbf{X}/k)*P(k\}/\Sigma\{P(i)*P(\mathbf{X}/i)\right\}$$

Fuzzy set theory can be also extended to clustering. Figure 11.8.2 shows an example of land cover classification using Fuzzy set theory. In this classification, the concrete structure (code 90), with clearly defined characteristics, was first classified using the ordinary maximum likelihood method, while the loosely defined urban classes were classified by the fuzzy based maximum likelihood method.

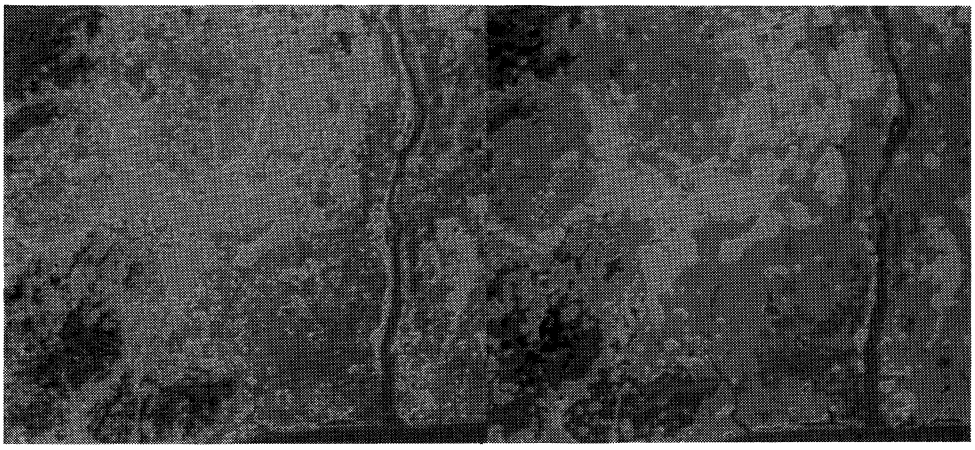


a) Classification of "child","young",and
"middle age" by the ordinary case of set theory



b) Classification of "child","young",and
"Middle age" by fuzzy method

Figure 11.8.1 Comparison between the ordinary case of set theory and fuzzy theory



a) First classification (code 90) by
the maximum likelihood method

b) Second classification by fuzzy
based maximum likelihood method

Figure 11.8.2 An example of land cover classification with Fuzzy theory

11.9 Classification using an Expert System

Experts interpret remote sensing images with knowledge based on experience. However computer assisted classification utilizes only very limited expert knowledge. The **expert system,** therefore, is a problem solving system which supports expert knowledge in a computer based system.

The following two types of knowledge are required for an expert system in remote sensing.

(1) Knowledge about image analysis
Procedures for image analysis can be made only with adequate knowledge about image processing and analysis. A feedback system should be introduced for checking and evaluating the objectives and the results.

(2) Knowledge about the objects to be analyzed
Knowledge about the objects to be recognized or classified should be introduced in addition to the ordinary classification method. The fact that forest does not exist over 3,000 meters above sea level, is one example of the type of knowledge that can be introduced.

Table 11.9.1 shows a list of knowledge required for delineating a tidal front in sea surface condition mapping. Figure 11.9.1 shows the sea surface condition map that was interpreted by an expert. Such knowledge will assure an increase in the accuracy or reliability of classification.

In many cases, knowledge can be represented as "if A is ..., then B becomes...." which is called the IF/THEN rule or production rule.

If the IF/THEN rule is fuzzy, then Fuzzy set theory can be also introduced to the expert system.

Figure 11.9.2 shows an example of the delineation of a tidal front using the expert system.

The expert system can be integrated with a geographic information system (GIS). It is necessary to accumulate experiences and to evaluate the knowledge for an expert system to be operationally applied.

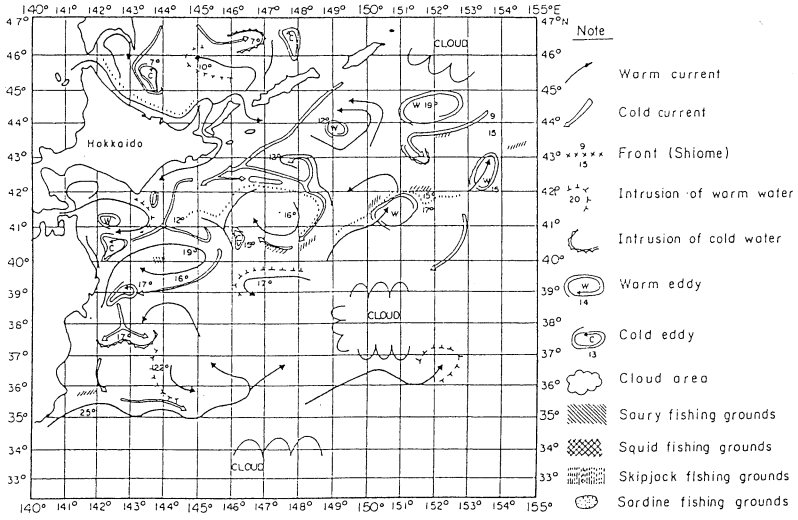
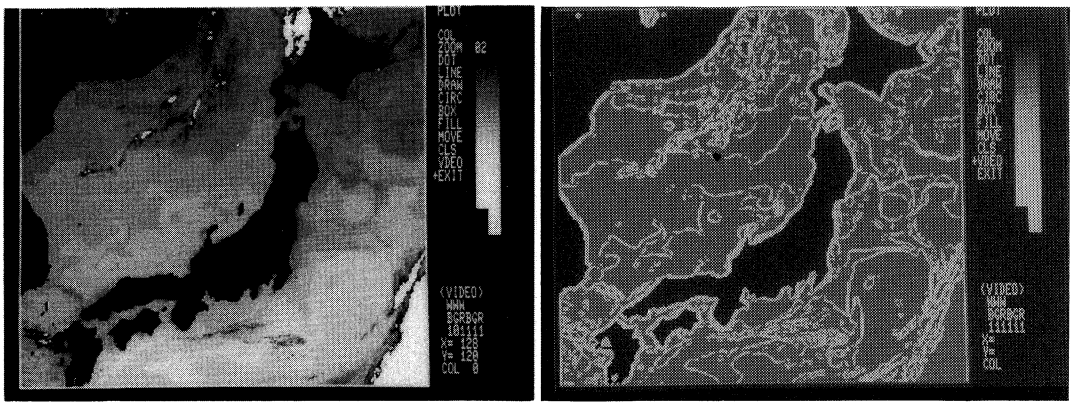


Figure 11.9.1 The sea surface condition map by an expert

Table 11.9.1 A list of knowledge for delineating tidal front in the sea surface condition mapping

1. Tidal front is (sea surface temperature)/(distance) > χ
2. Generally χ varies by areas
3. Definition of SST and distance depend on sea area
4. Distance is measured to the direction of normal vector of tidal front
5. the phenomenon is more likely on offing than coastal zone
6. Sometimes some parallel tidal fronts concentrates
7. Inclination of temperature is large at the area where some tidal fronts concentrates
8. Tidal front is a curve and intermittent
9. Tidal fronts never crosses
10. Tidal front agrees with the direction of isothermal line
11. Tidal fronts occur in some particular sea area
12. Approximately north-south tidal fronts more likely occur than east-west



a) Satellite data b) result of abstraction

Figure 11.9.2 Abstraction of tidal front by expert system