Proc. Indian Acad. Sci. (Engg. Sci.), Vol. 6, Pt. 2, June 1983, pp. 135–144.

Data processing in remote sensing

B L DEEKSHATULU and D S KAMAT*
National Remote Sensing Agency, Balanagar, Hyderabad 500037, India
* RSA, Space Applications Centre, Ahmedabad 380053, India

Abstract. A brief overview of pattern recognition and image processing with special emphasis on the first topic and its application to remote sensing is presented. Some of the recent areas of work in pattern recognition are also highlighted.

Keywords. Pattern recognition; image processing; classification; remote sensing; data processing.

1. Introduction

Modern remote-sensor systems collect vast quantities of data. Such a flow of data creates data-management problems which have their impact in data transmission, storage and retrieval, input and output, image processing and pattern recognition. The increasing spatial and spectral resolution of satellites aggravates this problem. For example LANDSAT-4 thematic mapper transmits nearly 10 times more information than LANDSAT-2 & 3 on a per unit area basis.

2. Data storage and retrieval

Remote-sensor technology appears to be overwhelmed by its own information explosion. For this information to be useful, it must be organised and stored in a manner that allows a convenient and orderly search. The information capacity of an image is an index for storage requirements. Image compression and compaction is necessary for storage. 'Compression' reduces the original image to a simpler image. Compaction schemes encode data in an efficient manner on the basis of statistical redundancy present in the data.

3. Image processing and pattern recognition

Image processing can be classified as follows: (a) image restoration (b) noise abatement (c) image enhancement (d) image analysis.

3.1 *Image restoration*

Image restoration involves the removal of systematic degradations due to detector/scanner/satellite systems.

3.2 *Noise abatement*

Some types of noise in an image are completely predictable or can be directly deleted. However most types of noises are statistical in character, and can be predicted only on a probabilistic basis. It is still possible to reduce the effects of such noise by properly-designed image processing operations.

3.3 *Image enhancement*

Image enhancement refers to such operations as the selective increase of contrast, delineation and for accentuation of edges and similar alterations.

3.4 *Image analysis*

Image analysis is the process of segmentation of the image using techniques like thresholding, edge detection, texture analysis, template matching and tracking (Rosenfeld & Kak 1976). These techniques make use of the statistical details like histograms, etc. The next logical step after segmentation is the image classification or pattern recognition.

3.5 *Pattern classification*

Pattern classification can be defined as the assignment of a point in the feature space (*e.g.* a remotely sensed 'pixel' characterized by its reflectances in different spectral bands) to a proper pattern class. The techniques used to solve pattern classification problems can be grouped into two general categories, namely, the decision theoretic (or statistical) and the syntactic categories. In the statistical approach, a set of features are extracted from the patterns and recognition is achieved by partitioning the feature space. In the syntactic approach each pattern class is characterised by several subpatterns and a relationship between these.

Another classification of pattern recognition techniques is supervised and un-supervised methods. In the supervised method certain number of training samples are available for each class; these are used to 'train' the classifier. The unsupervised method is akin to learning without a teacher. The decision theoretic methods can again be divided into parametric and non-parametric methods. In parametric methods each pattern class is characterized by a statistical distribution which in turn is dependent on certain number of 'parameters'. The non-parametric methods do not assume any such distribution.

3.5a *Non-parametric methods* (*Fu* 1980): (i) *Linear discriminant functions:*

$$\text{Let} \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_N \end{bmatrix}$$

be the feature vector. Let $\Omega_1, \Omega_2, \ldots \Omega_m$ be the m classes. The problem is to assign X to the proper Ω. A linear discriminant function is a linear combination of x_is. The decision boundary between the regions Ω_i and Ω_j is in the form of

$$D_i(x) - D_j(x) = \sum_{k=1}^{N} W_k x_k + W_{N+1} = 0,$$

where W_k's are constants determined from the training samples. A sample **x** is assigned to class W_i if

$$D_i(x) > D_j(x), \quad \forall j,\ j \neq i.$$

3.5b *Minimum distance classifier* One of the important types of linear classifiers is the minimum distance classifier. Here distances between sample points and prototype training samples are used for classification. Suppose that m reference vectors R_1, $R_2, \ldots, R_m$ are given for the m classes. The minimum distance classifier assigns the input sample X to class Ω_i if

$|\mathbf{X} - R_i|$ is the minimum, where $|\ \ |$ represents the distance defined as

$$|\mathbf{X} - R_i| = [(\mathbf{X} - R_i)^T (\mathbf{X} - R_i)]^{1/2}.$$

3.5c *Nearest neighbour classifier* Here there is a set of reference vectors for each class and the input sample is assigned to that class to which the nearest reference neighbour belongs. The nearest neighbour classifier has received considerable attention in the literature. It has been shown that for the infinite sample case the asymptotic error of the nearest neighbour classifier will not be greater than twice the Bayes' error.

3.5d *K-nearest neighbour rule* Here the input sample is assigned to that class to which a majority of its k neighbours belong.

3.5e *Edited nearest neighbour rule* Here the training reference vectors are themselves classified using the nearest neighbour rule and those vectors which are misclassified are removed from the training set.

3.5f *Polynomial discriminant functions* An rth order polynomial discriminant function can be expressed as

$$D_i(X) = W_{i1} f_1(\mathbf{X}) + W_{i2} f_2(\mathbf{X}) + \ldots + W_{iL} f_L(X) + W_{i,L+1}$$

where $f_j(X)$ is of the form

$$x_{k_1}^{n_1} x_{k_2}^{n_i} \ldots x_{k_n}^{n_i} \text{ for } k_1, \ldots, k_i = 1, \ldots, N \text{ and } n_1, n_2, \ldots, n_r = 0 \text{ and } 1.$$

3.5g *Training of linear classifiers* In a linear classifier the constants $W_1, \ldots, W_{N+1}$

have to be evaluated using the training samples. This process can be illustrated using a two-class classifier. Let

$$Y = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_N \\ 1 \end{bmatrix} = \begin{bmatrix} X \\ 1 \end{bmatrix},$$

when X is the feature vector. Let T_1 and T_2 be two classes. Then the problem is to find a vector W such that

$$Y^T W > 0 \quad \text{for} \quad Y \in T_1,$$

and

$$Y^T W < 0 \quad \text{for} \quad Y \in T_2.$$

If the output of the classifier is erroneous then a new W' is calculated as $W' = W + \alpha Y$ where α is a positive number.

3.5h *Parametric classifiers: Bayes' classifier* One of the most widely used parametric classifiers is the Bayes' classifier. In parametric classification, a probability density function is assumed and the parameters of that distribution are estimated. Let $x_i, \ldots . x_N$ be random variables where x_i is the noisy measurement of the ith feature. Let $p(x/\Omega_j)$ be the conditional probability density function of class j and $P(\Omega_j)$ is the *a priori* probability of class Ω_j. The task of the classifier is to assign the input sample such that the probability of misrecognition is minimized.

The Bayes' decision rule is that

$$\mathbf{X} \sim \Omega_i$$

if

$$P(\Omega_i)p(x|\Omega_i) \geqslant P(\Omega_j)p(x|\Omega_j), \ \forall j.$$

Assuming Gaussian distribution with mean vector M_i and covariance matrix K_i

$$p(X|\Omega_i) = \frac{i}{(2\pi)^{(N/2)}|K_i|^{1/2}} \exp\left[-\tfrac{1}{2}(X - M_i)^T K_i^{-1}(X - M_i)\right],$$

then the decision boundary between classes i and j becomes

$$\log \frac{p(\Omega_i)}{p(\Omega_j)} - \tfrac{1}{2}\left[(X - M_i)^T K_i^{-1}(X - M_i) - (X - M_j)^T K_j^{-1}(X - M_j)\right] = 0.$$

The above rule is also referred to as the maximum likelihood classification estimation rule (MLE). The MLE has been very popular for classifying remotely-sensed data. Even hardwired MLE processors have been built as part of data analysis systems.

3.6 *Sequential classification*

When the cost of feature measurements is taken into account and if the features are measured sequentially, it makes sense to use these features sequentially so that the average number of features used is minimal. Also the computational burden increases with the number of features. So sequential methods can be used in situations where there are a large number of features. (For example the multispectral scanner has 11 channels). Consider the 2-class problem. Wald's sequential probability ratio test (SPRT) computes at the nth stage the ratio

$$\lambda_n = p_n(\mathbf{X}|\Omega_1)/p_n(\mathbf{X}|\Omega_2)$$

where $p_n(\mathbf{X}/\Omega_i)$ is the conditional density for class i using the first n features. λ_n is compared with two stopping boundaries A and B. If $\lambda_n \geqslant A$, $\mathbf{X} \sim \Omega_1$, if $\lambda_n \leqslant B$; $\mathbf{X} \sim \Omega_2$. If neither of the inequalities is true, then the next feature is included and the process is repeated. The two stopping boundaries are related to the error probabilities;

$$A = (1 - l_{21})/l_{12},$$

and

$$B = l_{21}/(1 - l_{12}),$$

where l_{ij} is the probability of deciding $\mathbf{X} \sim \Omega_i$ when actually $\mathbf{X} \sim \Omega_j$. For the m class ($m > 2$) situation the generalized SPRT is used.

3.7 *Clustering* (Diday *et al* 1980)

The basic idea behind clustering is that, when a certain number of objects are scattered over a certain dimensional space, using a distance measure, it is possible to identify subgroups among these objects such that the objects belonging to a subgroup are 'closer' to themselves than to the members of other subgroups. These subgroups are called clusters. The distance measure used plays a central role in identifying the clusters. The following are some of the common distance measures used.

(i) Minkowsky metric

$$d(\mathbf{X}_i, \mathbf{X}_q) = \left[\sum_{j=1}^{n} |x_{ij} - x_{qj}|^{1/\lambda} \right]^{\lambda}$$

(ii) Quadratic metric:

$$d(\mathbf{X}_i, \mathbf{X}_q) = (\mathbf{X}_i - \mathbf{X}_q)^T Q (\mathbf{X}_i - \mathbf{X}_q)$$

where Q is a nxn positive definite matrix.

(iii) Mahalanobis metric

$$d_5(X_i, X_q) = (\det W)^{1/p} (X_i - X_q)^{T-1} W^{-1} (X_i - X_q)$$

where W is the covariance matrix.

3.7a *Binary distance measures*: (Fu 1980). If the features are binary (presence or absence) then the following distance measures are used

$$\text{Russel \& Rao } d_1(X_i, X_q) = a/(a+b+c+e),$$

$$\text{Jaccard \& Needham } d_2(X_i, X_q) = a/(a+b+c),$$

$$\text{Dice } d_3(X_i, X_q) = a/(2a+b+c),$$

$$\text{Sokal \& Sneath } d_4(X_i, X_q) = a/[a+2(b+c)],$$

where a = no. of occurrence of $x_{ij} = 1$ and $x_{qj} = 1$; b = no. of occurrence of $x_{ij} = 0$, and $x_{qj} = 1$; c = no. of occurrence of $x_{ij} = 1$, and $x_{qj} = 0$; e = no. of occurrence of $x_{ij} = 0$ and $x_{qj} = 0$.

A cluster P_s is said to be homogeneous if

$$\mathbf{X}_i, \mathbf{X}_j \in P_s \text{ and } \mathbf{X}_k \notin P_s,$$

$$\Rightarrow d(\mathbf{X}_i, \mathbf{X}_j) \leqslant d(\mathbf{X}_i, \mathbf{X}_k)\, 2d(\mathbf{X}_i, \mathbf{X}_j) \leqslant d(\mathbf{X}_j, \mathbf{X}_k).$$

The general idea in clustering is to have some representation of each of k clusters and from the knowledge of these, to identify the objects. The well-known ISODATA algorithm is explained by an example given below.

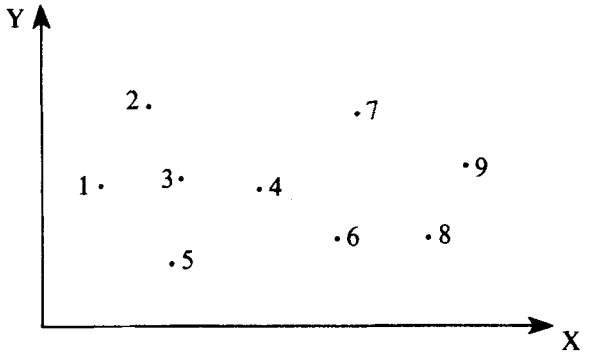

Figure 1. ISODATA example

Problem: Use ISODATA to classify the 9 samples (figure 1) into two clusters

The sample coordinates are

1. –(1·1, 2)
2. –(1·3, 2·7)
3. –(1·9, 2·3)
4. –(3, 2·2)
5. –(2·8, 1·2)
6. –(3·5, 1·5)
7. –(4·3, 2·5)
8. –(4·5, 1·6)
9. –(5·3, 2·4)

Assume points 5 and 6 as the potential cluster centres (seed points).

Assign points 1, 2, 3, 5 to point 5 since 5 is closer to these than 6

Similarly assign 4, 6, 7, 8, 9 to 6

New seed points: Centroid of 1, 2, 3, 5 and centroid of 4, 6, 7, 8, 9
These are (1·8, 2·05) and (4·1, 2·05) respectively.

Reassign the points to these seed points

1, 2, 3, 4, 5 are assigned to (1·8, 2·05) and 6, 7, 8, 9 are assigned to (4·1, 2·05)

The new seedpoints are centroids of (1, 2, 3, 4, 5) and (6, 7, 8, 9) = (2·02, 2·08), (4·4, 2)

The new assignments are (1, 2, 3, 4, 5) and (6, 7, 8, 9) which are the same as the previous assignments. So the two clusters are (1, 2, 3, 4, 5) and (6, 7, 8, 9).

The clustering scheme of Narendra & Goldberg (1977) starts first by computing a multi-dimensional histogram of the data. This is actually a data compression step because it tries to find out the distinct intensity vectors in the feature space. Therefore, instead of clustering all the available vectors in the feature space, only the distinct vectors are clustered. The frequency of occurrence of a particular intensity vector in the histogram is taken as density estimate and uses the valley seeking clustering algorithm of Koontze *et al* (1976) to cluster these distinct clusters.

Also the occurrences of frequencies can be used to obtain the stable maxima and stable minima in the distribution. Majumder *et al* (1981) have shown how to use these maxima and minima in defining the boundaries of clusters for one-dimensional data set.

3.8 *Per field classification and estimating mixture densities using dependent feature trees*

A recent development in classification is the use of fields (geographically contiguous pixels) rather than individual pixels. Here a group of pixels are classified by computing their density function and evaluating the distance between this function and the other class density functions. Here the underlying assumption is that all the pixels in that group (= field) belong to one class. If this assumption is valid then per field classification gives higher accuracy than per pixel classification.

A fundamental problem in unsupervised parametric classification is the estimation of the parameters (means, covariances etc. if the distribution is Gaussian) of the different classes from a combined mixture density of all the classes. Suppose the dimensionality of the data is n then for each class $n(n+3)/2$ parameters have to be estimated. Recently Chittineni (1982) has shown that if in each class we can assume that there is a tree-like relationship (*i.e.* each feature depends at most on one other feature) then the number of parameters to be estimated is only $(3n-1)$ per class. This result is a vast improvement and promises to be of far-reaching importance in estimating mixture densities.

3.9 *Learning with imperfectly labelled patterns*

Learning with an imperfect teacher has attracted considerable attention in the literature. Whitney & Dwyers (1966) obtained error bounds in a two-class situation on the performance of a nearest neighbour rule with an imperfect teacher. Kashyap & Blaydon (1966) proposed an iterative training procedure for a two-class case. Gimlin &

Ferrell (1974) studied the correction of labels using a nearest neighbour procedure. Chittineni (1979) considered the problem of learning with imperfectly labelled patterns. He considers the following model:

Let Ω and $\hat{\Omega}$ be the perfect and imperfect training set (a label is a class designate) labels, respectively, each of which takes values of 1, 2, ... M^a where M is the number of classes. Let $p(\Omega = i)$ and $p(X|\Omega = i)$ be the *a priori* and conditional densities of class i. $\beta_{ji} = P(\hat{\Omega} = i/\Omega = j), i, j = 1, 2, \ldots\ldots M$ are the probabilities of the imperfections. For the two-class case it can be shown that

$$P(\Omega = i) = \frac{1}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}}[\beta_{jj}P(\hat{\Omega} = i) - \beta_{ji}P(\hat{\Omega} = j)]; i, j = 1, 2.\ i \neq j.$$

and
$$p(\Omega = i|X) = \frac{1}{\beta_{11}\beta_{12} - \beta_{12}\beta_{21}}[\beta_{jj}p(\hat{\Omega} = i|X) - \beta_{ji}p(\hat{\Omega} = j|X)]\ i, j = 1, 2.$$
$$i \neq j$$

when the mislabelling is symmetric, the Bayes' error can be shown to be

$$P_i = \frac{1}{|2\beta - 1|}\hat{P}_i + \tfrac{1}{2}\left(1 - \frac{1}{|2\beta - 1|}\right),$$

where $\hat{P}_i$ is the Bayes' error with mislabelling.

3.10 *Texture*

Textural properties may play an increasingly important role in analysing remotely-sensed data. While visual texture is a difficult concept to define it is commonly said to involve the repetitive occurrence of local patterns in the given region (Rosenfeld *et al* 1980). Texture can be defined by describing local patterns and the rules of their arrangement.

Power spectrum analysis: The power spectrum gives information regarding the local patterns. The power spectrum at (x, y) is given by $|F(u, v)|^2 = F(u, v)\,F^*(u, v)$ where

$$F(u, v) = \int_{-\infty}^{\infty}\int \exp[-2\pi j(ux + vy)]f(x, y)\mathrm{d}x\mathrm{d}y$$
$$= \text{Fourier Transform of } f(x, y).$$

If the arrangement of local patterns over the region is periodic with period (u_0, v_0), then the power spectrum will have a high value at $(s/u_0, s/v_0)$ where s is the diameter of the region. Thus if the patterns are closely spaced the high values of the power spectrum will be spread out far from the region. Rosenfeld *et al* (1980) suggests that

$$F_1(r) = \int_0^{2\pi} |F(u, v)|^2 \mathrm{d}\theta \text{ and}$$

$$F_2(\theta) = \int_0^\infty |F(u, v)|^2 \mathrm{d}r$$

can be used as 'indices' of texture.

3.11 *Local property statistics*

The directional difference of averages taken over adjacent non-overlapping neighbourhoods is another index of texture. Let $A^r(x, y)$ denote the average of gray levels ($f(x, y)$'s) in a neighbourhood of radius r centred at (x, y); then a difference of non-overlapping A's in direction θ is defined by

$$D^{(r,\theta)}(x, y) = A^{(r)}(x + r\cos\theta, y + r\sin\theta) - A^{(r)}(x - r\cos\theta, y - r\sin\theta).$$

If the texture is fine-grained D will have high values for small values of r.

Joint gray level statistics: Joint frequency distribution of gray levels at various separations is another index of texture.

3.12 *Reduction in the number of bands*

The newer satellites have increasing number of spectral bands (for example LANDSATS 2 & 3 had 4 bands while LANDSAT-4 has 7 bands). While all these bands are required for different applications it is reasonable to suppose that a subset/combination of them would suffice for anyone field of application. The optimum way of selecting linear combinations of these bands so that the number of such combinations necessary to give a major percentage of the total information is much less than the total number of bands, is called the Karhunen–Loeve transformation or principal components transformation. This transformation involves the following steps:

Step 1: Obtain the covariance matrix of the pattern samples

Step 2: Obtain the eigenvalues and eigenvectors of this matrix

Step 3: Choose the $k(k < N$ = number of bands) eigenvectors corresponding to the k largest eigenvalues and form a transformation matrix.

Step 4: Transform the N-dimensional pattern vectors into k-dimensional vectors by using the transformation matrix.

In practice it is usually found that the first two principal components account for nearly 95% of the information.

4. State of Indian efforts in data processing in remote sensing

Indian scientists and engineers working in different academic institutions in India have made significant contributions in the fields of pattern recognition and image processing

in the last two decades. The Indian Institutes of Technology (IITs), Indian Institute of Science (IISc), Indian Standards Institution (ISI), etc., have a sound base in the theory and applications of pattern recognition techniques. The setting up of NRSA and the commencement of remote sensing activity at IIT, Bombay (in addition to the existing RSA at SAC) gave a fillip to this work. The supervised classification techniques like the maximum likelihood estimation and nonparametric techniques like ISODATA have become operational and are used in a routine manner in India. Techniques like fuzzy and syntactic classification are still in an experimental stage in this country. The position regarding data storage, creation of data bases, integration of different kinds of remotely-sensed data, etc., is not entirely satisfactory. Considerable work needs to be done in these areas in this country.

5. Conclusions

Digital data processing remote sensing has the advantage of speed and statistical analysis. While the standard processing includes error corrections and supervised classifications, special image processing has become a common feature for better image interpretation. Karhuan–Loeve transformation helps to a great extent the data reduction.

References

Chittineni C B 1979 *Pattern recognition image processing*, 52 IEEE Conf. (Chicago, IEEE)

Chittineni C B 1982 *Int. J. Remote Sensing* **3** 2

Diday E, Simon J C & Fu K S 1980 *Clustering analysis in digital pattern recognition* (New York: Springer Verlag)

Fu K S 1980 *Digital pattern recognition* (Springer Verlag)

Gimlin D R & Ferrell D R 1974 *IEEE Trans.* SMC-4 304–306

Kashyap R L & Blaydon C C 1966 *Proc. IEEE* **54** 1127

Koontze W L G 1981 *IEEE Trans. Comput.* C**25** 936

Majumder K L 1981 *Proc. machine processing of remotely-sensed data* Purdue Univ. LA R9 633

Manual of remote sensing (Am. Phys. Soc.) Vol. I

Narendra P M & Goldberg M 1977 *Pattern Recognition* **9** 207

Rosenfeld A & Kak A C 1976 *Digital picture processing* (New York: Academic Press)

Rosenfeld A, Weszka J S & Fu K S 1980 *Picture recognition in digital pattern recognition* (New York: Springer Verlag)

Whitney A W & Dwyers S J 1966 *Proc. of 4th Annual Allerton Conf. Circuit and System theory*, p. 96