PATTERN RECOGNITION OF REMOTELY SENSED DATA

Robert M. Haralick
Department of Electrical Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

I. Introduction

For us, Pattern Recognition refers to the automatic machine determination of salient patterns in remotely sensed image data. From the pattern recognition perspective, the world to be sensed is composed of units defined by the sensor. For digital imaging sensors, as a first approximation, the units can be thought of as small non-overlapping areas on the ground: one such area for each picture element (pixel) in the image. The sensor makes an ordered set of measurements on each unit sensed. The ordered set of measurements is called a measurement vector or measurement pattern. Each value measured in this set is a number proportional to the energy received by the sensor in some band of the electromagnetic spectrum at some specified observation time. The basic pattern recognition problem is first to automatically and consistently determine the informational class or category of each distinct region on the ground using the set of sensor measurement patterns and second to estimate the error rate for the automatically determined assignments.

Specific examples of pattern recognition for remote sensing applications include determining

(1) tree species composition in a forest
(2) hot spots of incipient forest fires
(3) natural vegetation cover types
(4) crop types
(5) State of health or stress vegetation
(6) percent of sedimentation in a river or lake
(7) percent of pollutant in a river or lake
(8) geological formation and rock types
(9) lineament patterns
(10) degree of mineralization
(11) number of small objects in a smooth background
(12) urban land use patterns

The automation of these tasks requires a corresponding variety of methods and techniques varying from simple to highly complex. It is

NATO ASI Series, Vol. F4
Pictorial Data Analysis. Edited by R.M. Haralick

the purpose of this chapter to describe the most commonly used techniques.

Books describing the principles of pattern recognition have been written by Sebestyen (1962), Nilsson (1965), Arkadev and Braverman (1966), Fu (1968), Kanal (Ed.)(1968), Watanabe (Ed.) (1969), Mendel and Fu (1970), Fu (Ed.)(1971), Andrews (1972), Fukunaga (1972), Meisel (1972), Patrick (1972), Watanabe (Ed.) (1972), Chen (1973), Duda and Hart (1973), Ullman (1973), Tou and Gonzalez (1974), Batchelor (1974), Young and Calvert (1974), Fu and Whinston (Ed.)(1977), and Batchelor (1978). Some of these books have been reviewed and the reader might be interested in consulting the reviews listed in the table before attempting to read any of these books.

Shorter reports and review articles include those by Nagy (1968), Ho and Aggarwala (1968), Fu, Landgrebe, and Philips (1969), Casy and Nagy (1971), Nagy (1972), Kanal (1972) and Kanal (1974). Reprints of important pattern recognition articles can be found in Sklansky (1973) and Aggarwala (1976). The May 1979 issue of the IEEE Proceedings was a special issue on pattern recgonition and image processing. Journal papers on pattern recognition appear in the IEEE Transactions on Systems, Man and Cybernetics, and IEEE Transactions on Pattern Analysis and Machine Intelligence. The Pattern Recognition Society publishes a journal called Pattern Recognition. Conference papers appear in the International Joint Conference on Pattern Recognition, The Pattern Recognition and Image Processing Conference, The Purdue Symposium on Machine Processing of Remotely Sensed Data, and the Environmental Research Institute of Michigan Remote Sensing of Environment Conferences.

Harry Andrew	Introduction to Mathematical Techniques in Pattern Recognition, Prentice Hall, New Jersey, 1972, 504 pages.	IEEE Information Theory IT-19 No. 6, November, 1973, p. 831.
Richard Duda and Peter Hart	Pattern Classification & Scene Analysis, Wiley, New York, 1973,482 pages.	IEEE Computer Transacaction, Vol. C-23, No.2, February. IEEE Information Theory Vol. IT-19, No.6, November 1973, p. 827-829.
King-Sun Fu	Syntactic Methods in Pattern	IEEE Systems Man Cyber-

	Recognition, Academic Press, New York, 1974, 397 pages.	netics, Vol. SMC6, No. 8, August, 1976, p. 590.
Keinosuke Fukunaga	Introduction to Statistical Pattern Recognition, Academic Press, New York, 1972, 382 pages.	IEEE Systems Man Cybernetics, Vol. SMC-4, No. 2, March, 1974, p. 238. IEEE Information Theory Vol. IT-19, No.6, November, 1973, p. 827-829.
William Meisel	Computer-Oriented Approaches to Pattern Recognition, Academic Press, New York, 1972, 262 pages.	IEEE Systems Man Cybernetics, Vol. SMC-3, No.2, March, 1973, p. 209. IEEE Computer Transactions, Vol. C-23, No.1, January, 1974, p. 112. IEEE Computer Transactions, Vol. C-22, No.4, April, 1973, p. 429. IEEE Information Theory, Vol. IT-19, No. 6, November,1973, pp. 832-833.
Edward Patrick	Fundamentals of Pattern Recognition, Prentice Hall, New Jersey, 1979, 528 pages.	IEEE Systems Man Cybernetics, Vol. SMC-3, No.5, September,1973, p. 528. IEEE Information Theory Vol. IT19, No.6, November, 1973, pp. 830-831.
Julius Tou and Rafael Gonzales	Pattern Recognition Principles, Addison-Wesley, Mass. 1974, 377 pages.	IEEE Systems Man Cybernetics, Vol. SMC-6, No.4, April,1976, pp. 632-633.

Julian Ullman	Pattern Recognition Techniques, Crane-Russak, New York, 1973, 412 pages.	IEEE Computer Transactions, Vol. C23, No.2, February,1974, pp. 220-222. IEEE Information Theory, Vol. IT-20, No.3, May, 1974, p. 400.
Satosi Watanabe (Ed.)	Methodologies of Pattern Recognition, Academic Press, New York, 1969, 579 pages.	IEEE Information Theory Vol. IT-17, No.5, Sept. 1971, pp. 633-634.

Table 1 lists various books on statistical pattern recognition and where they have been reviewed

II. Summary

To do the pattern recognition automation job, we must define the class of entities of interest, that is, between which kinds of objects we must discriminate; we must choose instruments or sensors which can measure the environment in which the objects occur; we must provide a methodology permitting the recognition of an object in the class of objects of interest from those not in the class of objects of interest; and using this methodology we must construct a decision rule which will decide what kind of object a particular object is, on the basis of the measurements made from the observed small area ground patches.

Defining the class of objects of interest should be easy since it is an intrinsic part of the automation need. We will see however, that it is not so easy since the sensor may not gather sufficient information to allow the discrimination to take place. In these cases we will prefer to define our classes to be the more discriminable ones even though they may be less interesting to us. To help us do this we need to employ a clustering process which tells us what are the naturally distinguishable classes given the sensor's data.

Choosing the measuring instruments or sensors and designing a way to preprocess - to standardize, to normalize, and to extract the revelant information in its simplest form from the measurements - so that objects of interest can be simply recognized from those of non-interest and so that each class or category of objects of interest has a particularly simple description in terms of the preprocessed measurements are amoung the most difficult problems in the pattern recognition area. These problems are called feature-extraction or preprocessing problems and are concerned with presenting in some standard form only the simplest most important information to the decision rule.

Finally, the problem in constructing a decision rule we call the pattern discrimination problem. It is based on a probability model and it allows us to estimate the error rates of the automatic decision process.

Most pattern recognition of remotely sensed image data is done processing each pixel's information separately or independently. This means that a category assignment is made to each pixel purely on the basis of its own information. Processing proceeds on a pixel by pixel basis over the entire image.

When the pixel's information consists only of the sensor measurement pattern obtained from one observation time, the measurement pattern is called a multispectral feature vector and the kind of pattern recognition is called multispectral pattern recognition. When spectral information from more than one observation time for the same ground area are stacked in the same measurement pattern vector, this kind of pattern recognition is called multispectral multitemporal pattern recognition. When the measurement pattern for each pixel contains spectral information from its associated ground area as well a neighboring ground area or when the decision rule which makes category assignments uses the information from a pixel and some of its neighboring pixels, the pattern recognition is called spatial pattern recognition.

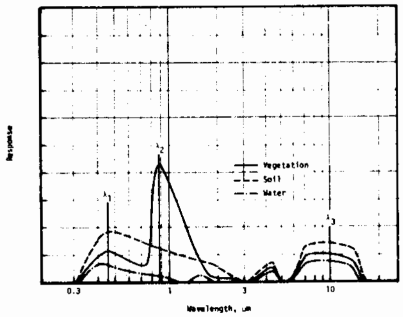

Figure 1. Typical relative response curves for different materials, illustrating the possibility of discrimination by comparision of curves at different wavelengths. Source: Landgrebe (1972).

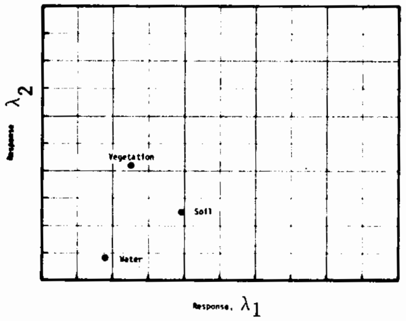

Figure 2. The categories vegetation, soil, and water distinct responses on wavelengths λ_1 and λ_2 Shown in this figure are these categories plotted in a measurement space whose axes are their λ_1 and λ_2 responses. Source: Landgrebe (1972).

III. Principles of Spectral Discrimination

In order to understand the pattern discrimination methodology consider a simplified example. Suppose that there are three types of surface cover material: vegetation, soil, and water. Suppose that each of these has a unique spectral response which does not vary with season, atmosphere haze, sunangle etc. Let these be the responses shown in figure 1. Now select two wavelengths λ_1 and λ_2 for a remote sensor to make some measurements. Then, for each surface cover category, use wavelengths λ_1 to determine its spectral measurement pattern. Plot these in measurement space as shown in figure 2. Since they obviously plot nicely separated from each other we would expect no difficulty in designing a decision rule to recognize these categories. Anytime a new meaurement pattern needs to be assigned a category we see if it lies as the point in measurement space associated with vegetation, or soil, or water. If it does, we assign it the corresponding category. If it doesn't we assign it unknown.

In reality, the spectral response patterns from these surface categories as well as others vary due to natural random variations, systematic seasonal causes, and atmospheric haze, etc. There is not a unique measurement pattern associated with each category. Rather, associated with each category is a probability distribution indicating for any measurement pattern the relative frequency of occurrence that it may arise from a ground area of the given category.

If, using some training data, we plotted five observations of each of three vegetation categories, soybean, corn and wheat, we might obtain the measurement space plot of figure 3. To assign a new measurement pattern v to one of the classes is now not such an easy problem. In essence we must use our training observations to estimate for each new measurement pattern v, the probability that soybeans, corn or wheat is its true category. If we can do that we can associate with each measurement that category having highest conditional probability given the measurement. In effect, this association partitions measurement space as shown in figure 4. Since our new measurement pattern is in the part of measurement space associated with soybeans the decision rule assigns it to the soybean class.

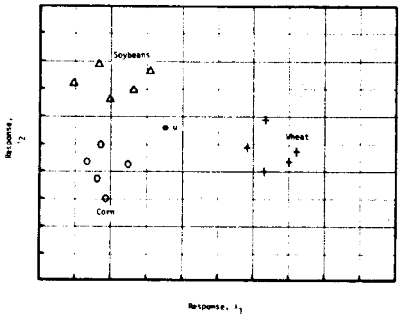

Figure 3. A given material will not always have the exactly same response in a group of samples but each material tends to cluster together. A typical two-dimensional sampling of three materials is shown. Source: Landgrebe (1972).

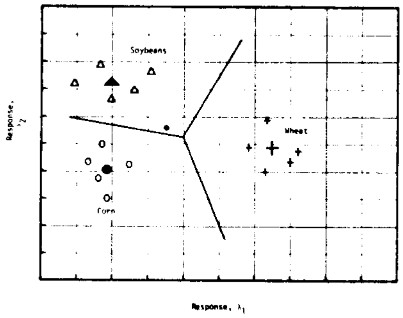

Figure 4. Division of two-dimensional sampling space into domains assigned to different materials. In this case the unknown point u is considered to be soybean because of its location in the sampling sapce. Source: Landgrebe (1972).

The procedure by which the measurement space of figure 3 was partitioned is simple. Use the training data for each class to determine the class sample mean. Then partition measurement space is that each class has associated with it all the measurement patterns closest to its sample mean. Unfortunately, without a probability model we cannot say that this procedure is the one that yields the lowers error rate or maximizes any utility function. Although, there is a probability model under which this is the appropriate thing to do.

It is the purpose, therefore, of the next sections to develop a probabilistic decision theoretic model for pattern discrimination which suggests techniques for decision rule construction having certain optimal properties which we can measure in terms of utility or economic consequences.

IV. Economic Consequences of Decisions

For each pattern d belonging to D, $d \varepsilon D$, a decision rule f assigns a category alternative c^k from the set of category alternatives $C = \{ c^1, \ldots\ldots, c^K \}$. The assignment may be deterministic or probabilistic. In any case, we assume that the assignment by the decision rule of category alternative c to a pattern d measured from a unit u caries economic consequences. These economic consequences are determined by the people who need to automate the discrimination ability of the trained human observer. The consequences are generally good when the chosen category alternative c is in fact the true category indentification of the unit u. The consequences are generally bad when the category alternative c is not the true category identification of the unit u. Because such identification decisions must be made, and because they cause consequences when they are made, we may view the goal of decision rule construction as the construction of a decision rule which in some sense maximizes the good consequences.

To speak of maximizing good consequences implies that we must have some numerical measure indicating the economic gain or loss of the consequence when the decision rule assigns category c^i to a unit u with measurement d when the true category identification of unit u is category c^j. Let $e(c^j,c^i)$ be the net worth or economic gain of such a consequence. In general, $e(c^i,c^i)$ will be positive signifying a gain for a correct identification, and $e(c^i,c^j)$, for $i \neq j$, will be negative signifying a loss for an incorrect identification.

In determining a decision rule, we must choose a criterion of optimality by which we can judge the worth of the decision rule on the basis of the various economic gains or losses of the consequence (c^i,c^j). The optimality criterion defines how to judge how well the decision rule balances, in terms of these gains and losses, the possible consequences of its decision. The most often used criterion is one which defines the best decision rule to be one which maximizes the expected gain under certain given conditions. Such a rule is called a Bayes decision rule.

Let us consider the economic gains of the possible consequences given that a unit u has measurements d. These gains are illustrated simply in Figure 5. Suppose the decision rule assigns a unit u having measurements d to category c^i This asignment, at best, however, is only an educated guess; the true category identification for unit u can actually be one of $c^1, c^2, \ldots, c^K$. In Figure 5 the decision rule assignment of c^i corresponds to a selection of the i^{th} column. The true category identification of unit u corresponds to a selection of some row. This row intersected with the i^{th} column yields an entry which is the economic gain consequence.

	c^1	c^2		c^K
$P_d(c^1)$	$e(c^1, c^1)$			$e(c^1, c^K)$
$P_d(c^2)$				
N A T U R E				
$P_d(c^K)$	$e(c^K, c^1)$			$e(c^K, c^K)$

Figure 5 shows the economic gains obtained under various alternatives conditioned on the measurement d being made of a unit K. Given that the observed measurement is d, the probability that nature choose category c^j, corresponding to the j^{th} row, is P_d (c^j). The decision rule will choose some category c^k, corresponding to the k^{th} column. the result of nature choosing category c^j and the decision rule choosing category category c^k is the economic consequence (c^j, c^k).

The question of concern is how often will the true category identification of a unit u be category c^j when the unit u has measurement d. We denote by $P_d(c^j)$ the probability of the true category identification of a unit u being in category c^j given that the unit u has measurements d. It is these conditional probabilities which can be estimated from the training data or ground truths data.

The decision rule has no information regarding the true category identification of any unit. It only knows that the unit gives rise to a pattern d and it has available estimates of the conditional probabilities $P_d(c^k)$, $k = 1,2,\ldots,K$. The decision rule must assign the unit to a category, say c^i This corresponds to a selection of the i^{th} column. For this course of action a number of difference consequences can occur. If the true category identification is c^1, then the gain of the consequence (c^1, c^i) is $e(c^1,c^i)$. If the true category identification is c^2, then the gain of the consequence (c^2,c^i) is $e(c^2,c^i)$. The next section discusses a decision rule construction procedure which maximizes the expected gain.

V. The Bayes Decision Rule Maximizes Expected Gain

Let $f_d(c)$ denote the probability that the decision rule assigns the category c to the unit given that the unit has pattern measurement d. Since for any pattern d, there is no reason to suppose any interaction or collaboration between nature, who may be thought of as choosing the true category identification, and the pattern discriminator, which may be thought of an employing the decision rule to assign categories we may assume that nature and the pattern discrimination are statistically independent. Thus, the probability that the unit has measurements d and the decision rule assigns the category c^k to the unit and the true category identification for the unit is c^j may be written as $f_d(c^k)P_d(c^j)$ $P(d)$. Therefore, the expected gain for the decision rule f may be expressed by

$$E[e;f] = \sum_{d\varepsilon D} \sum_{j=1}^{K} \sum_{k=1}^{K} e(c^j,c^k) f_d(c^k)\; c({}^j)P(d).$$

To see how to find that decision rule which maximizes the expected gain, we rewrite the expression for E[e;f] as

$$E[e;f] = \sum_{d\varepsilon D} P(d) \sum_{k=1}^{K} f_d(c^k) \sum_{j=1}^{K} e(c^j,c^k)\; P_d(c^j).$$

P(d), being the probability of measuring pattern d for a unit, is non-negative. Hence E[e;f] will be maximized (maximum taken over all f) if and only if for each $d\varepsilon D$ the expected gain given d using f is maximized ; that is,

$$E[e|d;f] = \sum_{k=1}^{K} f_d(c^k) \sum_{j=1}^{K} e(c^j,c^k)P_d\ (c^j) \text{ is maximized.}$$

Since $\sum_{k=1}^{K} f_d(c^k) = 1$ and $f_d(c^k) \geq 0$, $k = 1, 2,\ldots,k$, it is easy to see that the maximum of the above expression is

$$\max_{\substack{i \\ i = 1,2,\ldots,K}} \sum_{j=1}^{K} e(c^j,c^i)P_d(c^j)$$

and the decision rule f will certainly achieve this maximum if

$$f_d(c^i) = \left.\begin{matrix} 1, i = k \\ 0, i \neq k \end{matrix}\right\} \text{ where k is any index such that}$$

$$\sum_{j=1}^{K} e(c^j,c^k)P_d(c^j) \geq \sum_{j=1}^{K} e(c^j,c^i)\ P_d(c^j),\ i=1,2,\ldots,K.$$

In this case the optimal decision rule can be deterministic if the index k is unique or it can be either deterministic or probabilistic if k is not unique. Any optimal decision rule is called a Bayes rule.

For example, suppose there are three categories c^1, c^2, and c^3 with conditional probabilities and economic gains for the various alternatives and consequences shown in Figure 6. The optimal decision rule will assign the unit u to category c^3 since the average gain for now is 5/6 which is larger than the average gain for row 1 which is -1/3 or for row 2 which is 1/2.

	c^1	c^2	c^3
$P_d(c^1)=1/6\ c^1$	4	-2	0
$P_d(c^2)=1/2\ c^3$	0	1	0
$P_d(c^3)=1/3\ c^3$	-1	0	3

Figure 6. Illustrates the economic gains for an example problem where the pattern measurements d are made on a unit and there are three possible categories.

VI. Bayes Decision Rules and Category Prior Probabilities

It is often the case that the conditional probabilities P(c|d) are not known but that the conditional probabilities P(d|c) of the measurements given the categories are known. Fortunately, there is a well known relationship between P(c|d) and P(d|c) which involves the prior probabilities of P(d) and P(c) of the measurements and categories, respectively.

By the definition of conditional probability , we may express $P_d(c)$ by

$$P_d(c) = \frac{P_c(d)P(c)}{P(d)}$$

so that the average gain obtained by the use of decision rule f may be rewriten as

$$E[e;f\} = \sum_{d \varepsilon D} \sum_{k=1}^{K} \sum_{j=1}^{K} f_d(c^k)e(c^j,c^k)P_{c^j}(d)P(c^j).$$

E[e;f] is maximized if and only if for each dεD, the gain conditioned on d,

$$E[e|d;f] = \sum_{j=1}^{K} f_d(c^k) \sum_{j=1}^{K} e(c^j,c^k)P_{c^j}(d)P(c^j)$$

is maximized. The maximum value of E[e|d;f] is

$$\sum_{j=1}^{K} e(c^j,c^k)P_{c^j}(d)P(c^j)$$

where k is some index for which

$$\sum_{j=1}^{K} e(c^j,c^k)P_{c^j}(d)P(c^j) \geq \sum_{j=1}^{K} e(c^j,c^i)P_{c^j}(d) \; P(c^j), i=1,2,\ldots,K.$$

An optimal deterministic decision rule f may therefore be defined by

$$f_d(c^i) = \left. \begin{array}{l} 1, i = k \\ 0, i \neq k \end{array} \right\} \text{ where k is any index such that}$$

$$\sum_{j=1}^{K} e(c^j,c^k)P_{c^j}(d)P(c^j) \geq \sum_{j=1}^{K} e(c^j,c^i)P_{c^j}(d) \; P(c^j), i=1,2,\ldots,K.$$

Note the strong dependence which f has on the category probalitity P(c). Because of this, anytime we define an optimal Bayes decision rule we must state that it is optimal only relative to the category prior probability function P(c).

VII. Maximin Decision Rule

Figure 7 illustrates the expected gain of a Bayes decision rule in a two-category classification problem with the identity gain function. Selecting a value of prior probability, the corresponding value of expected gain is the highest expected gain achievable by any decision rule. Therefore, use of any decision rule which is not a Bayes rule is guaranteed to perform below the curve. In particular, if a Bayes rule is used in a new situation where the encountered prior probability function differs from the one employed in the design, then the Bayes rule is not optimal in the new situation.

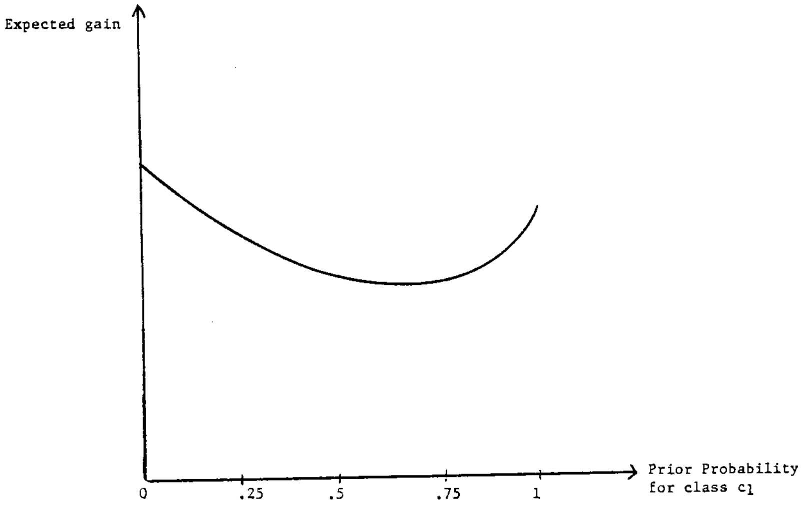

Figure 7 illustrates how the expected gain of a Bayes decision rule can vary with a change in prior probability for class c_1 in some two class example problem. Notice that as the prior probability for class c_1 becomes 1, the prior certainty reflects itself in an a posteriori certainty which makes the expected gain high. When the prior probability for class c_1 becomes 0, the prior probability for class c_2 becomes 1 and the situation is similar. For class c_1 prior probabilities between 1 and 0. the prior situation is less certain and the expected gain must be less than the end cases. The shape of the function is guaranteed to be convex.

Recognizing this, a conservative decision rule decision designer will attempt to construct a decision rule which maximizes the smallest gain achieved by the decision rule under some encountered prior probability function. It turns out that the expected gain for a decision rule which maximizes the smallest gain as the encountered prior probability function varies has a value equal to the smallest possible Bayes gain (the lowest point on the curve of Figure 7.) This value is the same regardless of the actually encountered prior probability function. This kind of decision rule is called a maximin decision rule since it maximizes the minimum expected gain. In general, it is not a deterministic decision rule and designing a maximum decision rule is equivalent to solving a large linear programming problem.

VIII. The Gaussian Assumption

The conditional probability $P_c(d)$ of the data measurement vector d given the category c plays an essential role in decision rule determination. $P_c(d)$ could be stored as a table. However, because of the large number of possible data vectors, $P_c(d)$ is often represented as parametric function, the parameters being the category mean mesurement μ_c and its covariance matrix $\sum_c$. The simplest probability density function having these parameters is the Gaussian one which is defined by

$$P_c(d) = \frac{1}{(2\pi)^{N/2} |\sum_c|^{1/2}} e^{-1/2(d-\mu_c)' \sum_c^{-1} (d-\mu_c)}$$

where N is the dimension of the measurement vector d.

In the case of identity gain function e and equal probabilities for the prior P(c) for each category c, the Bayes decision rule assigns measurement d to that category c minimizing

$$\log |\sum_c| + (d-\mu_c)'\sum_c^{-1}(d-\mu_c)$$

This kind of decision rule is sometimes called quadratic or piecewise quadratic because the decision boundaries they form in measurement space are piecewise hyperquadratic boundaries.

In the case that the covariance matrices for all categories are equal, the Bayes decision rule reduces to assigning the measurement d to that category c minimizing the Mahalanobis distance

$$(d-\mu_c)' \sum^{-1}(d-\mu_c)$$

between the measurement vector d and the category mean vector μ_c for category c. This decision does simplify to a linear decision rule. Assign measurement d to that category c maximizing

$$\left[\mu ' \sum^{1}\right] d - \left[\frac{\mu_c'\sum^{-1}\mu_c}{2}\right]$$

By precomputing the terms in parentheses, the number of multiply and add operations for this decision rule is only N + 1 per category, a significant saving over the quadratic rule, especially when the dimensionality N is large.

IX. Feature Selection

Multitemporal multispectral remotely sensed imagery can produce a ten or twenty dimensional data vector for each pixel. The data has inherent redundancies and processing all of it or storing all of it may not be cost effective. Feature selection procedures are used to select those dimensions most suitable for processing.

There are two kinds of feature selections depending on whether the classses and their statistics are known or not known. If they are not known, the best feature selection procedure is called principle components. If they are known, the easiest to use feature selection is based on Bhattacharyya distance.

1. <u>Principal Components</u>

Principal components is a standard statistical technique for selecting that subspace of given dimension in which the most data variance lies. If $x_1, \ldots, x_n$ are the sample data vectors, μ the sample mean vector, and $\sum$ the sample covariance matrix, the best K dimension in which to project the data would be that K-dimensional subspace spanned by the K eigenvectors of $\sum$ having largest eigenvalues. Thus if T is a matrix whose K rows are these eigenvectors, the K principal components of $x_1, \ldots, x_n$ is $Tx_1, \ldots, Tx_n$, each Tx_n being a K-dimensional vector.

2. <u>Bhattacharyya Distance</u>

The Bhattacharyya distance is a measure of the separability between two classes. For two Gaussian classes having means and covariances μ_1, $\sum_1$ and μ_2, $\sum_2$ respectively, the Bhattacharyya distance is given by

$$1/8\ (\mu_1 - \mu_2)' \left[\frac{\sum_1 + \sum_2}{2}\right] (\mu_1 - \mu_2)$$

$$+\ 1/2\ \ln \left\{ \frac{\dfrac{|\sum_1 \sum_2|}{2}}{|\sum_1|^{1/2} |\sum_2|^{1/2}} \right\}$$

To use this distance measure for selecting the best K features from the original N dimensions on an L class problem, the Bhattacharyya distance needs to be calculated between each of the L(L-1)/2 pairs of classes for each of the $\begin{bmatrix} K \\ N \end{bmatrix}$ possible ways of choosing K features from N dimensions. The K dimensions which are best are those K dimensions whose sum of the Bhattacharyya distances between the L(L-1)/2 pairs of classes is highest. The Bhattacharyya distance between a pair of classes for a selection of K dimensions is calculated using the mean and covariance matrix in the selected K dimensions.

REFERENCES

1. A. K. Aggrawala, (Ed.). Machine Recognition of Patterns, IEEE Press, New York, 1977.

2. Harry Andrews, Introduction to Mathematical Techniques in Pattern Recognition, Prentice Hall, New Jersey, 1972, 504 pages.

3. A. G. Arkadev and E. M. Braverman, Computer and Pattern Recognition, Thompson Book Company Inc., Washington, D.C., 1966.

4. B. G. Batchelor, (Ed.), Pattern Recognition Ideas in Practice, Plenum Press, New York, 1978.

5. B. G. Batchelor, (Ed.), Pattern Approaches to Pattern Classification, Plenum Press, New York, 1974.

6. Casey and Nagy, "Advances is Pattern Recognition" Scientific American, Vol. 224, No. 4, April 1971, p. 56-71.

7. C. H. Chen, Statistical Pattern Recognition, Hayden, Washington, D.C., 1973.

8. Richard Duda and Peter Hart, Pattern Classification & Scene Analysis, Wiley, New York, 1973, 482 Pages.

9. K. S. Fu, D. A. Landgrebe and T. L. Philips, "Information Processing of Remotely Sensed Agricultrual Data," Proceedings of the IEEE, Vol. 57, No. 4, April 1969, pp. 639-653.

10. K. S. Fu, Syntactic Methods in Pattern Recognition, Academic Press, New York, 1974, 397 Pages.

11. K. S. Fu, (Ed.), Pattern Recognition and Machine Learning, Plenum Press, New York, 1971.

12. K. S. Fu and A. B. Whinston, Pattern Recognition Theory and Application, Noordhoff-Leyden, Netherlands, 1977.

13. K. S. Fu, Sequential Methods in Pattern Recognition and Machine Learning, Academic Press, New York, 1978.

14. Keinosuke Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, New York, 1972, 382 pages.

15. Y. C. Ho and A. K. Aggrawala, "On Pattern Classification Algorithms Introduction and Survey," Proceedings of the IEEE, Vol. 56, No. 12, December 1968, pp.2101-2114.

16. L. N. Kanal, "Interactive Pattern Analysis and Classification System: A Survey and Commentary," Proceedings of the IEEE, Vol. 60, No. 11, October 1972, pp. 1200-1215.

17. L. N. Kanal, "Patterns in Pattern Recognition: 1968-1974," IEEE Transactions on Information Theory, Vol. IT-20, No. 6, November 1974, pp. 697-722.

18. L. N. Kanal, (Ed.), Pattern Recognition, Thompson Book Company, Washington, D.C., 1968.

19. William Meisel, Computer-Oriented Approaches to Pattern Recognition, Academic Press, New York, 1972, 262 Pages.

20. J. M. Mendel and K. S. Fu, (Eds.), Adaptive, Learning and Pattern Recognition Systems: Theory and Applications, Academic Press, New York, 1970.

21. George Nagy "State of the Art in Pattern Recognition" Proceedings of the IEEE, Vol. 56, No.5, May 1968, pp. 836-862.

22. George Nagy "Digital Image-Processing Activities in Remote Sensing for Earth Resources" Proceeding of the IEEE, Vol. 60, No. 10, October 1972, pp. 1177-1199.

23. N. J. Nilsson, Learning Machine, McGraw-Hill, New York, 1965.

24. Edward Patrick, Fundamentals of Pattern Recogintion,Prentice hall, New Jersey, 1972, 528 pages.

25. G. S. Sebestyen, Decision Making Processes In Pattern Recognition, The Macmillan Company, New York, 1962.

26. J. Sklansky, (Ed.), Pattern Recognition, Introduction and Foundation, Dowden Hutchinson and Rose, Inc., Pennsylvania 1972.

27. Julius Tou and Rafael Gonzalez, Pattern Recognition Principles, Addison-Wesley, Mass. 1974, 377 pages.

28. Jullian Ullman, Pattern Recognition Techniques, Crane-Russak, New York, 1973, 412 pages.

29. Satosi Watanabe, (Ed.), Methodologies of Pattern Recognition, Academic Press, New York, 1969, 579 pages.

30. Satosi Watanabe, Frontiers of Pattern Recognition, Academic Press, New York, 1972.

31. T. Y. Young and T. W. Calvert, Classification, Estimation, and Pattern Recognition, American Elsevier, New York, 1974.