LARS Information Note 101866

On Pattern Recognition

by
G. P. Cardillo
D. A. Landgrebe

The Laboratory for Applications of Remote Sensing

Purdue University, West Lafayette, Indiana

LARS INFORMATION NOTE 101866

Purdue University

ON PATTERN RECOGNITION*

Introduction

This short introductory report is intented to familiarize the reader, in non-mathematical terms, with the field of automatic pattern recognition and its applications. We begin with a discussion and some examples typifying the problems included within the realm of pattern recognition. The basic model of a pattern recognition device consisting of a receptor and a categorizer is discussed, as are a number of approaches to the solution of pattern recognition problems. From this discussion it will be seen that although the fields of application of pattern recognition are very diverse, important properties of each problem can be extracted and formulated within a common framework.

*This Information Note was prepared by G. P. Cardillo and D. Landgrebe of the Purdue data group.

What is Automatic Pattern Recognition

Generally, the term pattern recognition (PR) as used in the technical literature refers to the development of techniques and equipment for the automatic recognition of patterns. The emphasis here is on automatic, since this field has been developed to handle problems in which the large quantity of data demands complete reliance upon a machine for classification.

There appear to be similarities between PR and photo interpretive techniques (PI). As with PI, PR requires the development of a key, a set of tests which are to be carried out on a candidate pattern to determine its correct classification. The similarity ends at this point, however, due to the nature of the sets of tests in the two cases and the way they must be implemented. In the case of PI, the tests are usually relatively sophisticated and require human attention. On the other hand, the purpose of PR is to permit the complete removal of man from the process in order to be able to process data faster.

Thus in comparing PR and PI, it may be said that PI is generally more suited for problems of higher sophistication involving lower data quantities, while just the reverse is true of PR.

In order to further clarify what is meant by the terms pattern and pattern recognition, a number of examples of current and important problems are presented.

Probably the first thing that comes to mind upon hearing the term pattern recognition is the problem of automatically recognizing various geometrical patterns. Examples of this type of problem are:

1) Reading of typed, printed, or handwritten text

2) Recognition of a person from his handwriting

3) Distinguishing manmade from natural objects on aerial photographs

But pattern recognition is by no means limited to these cases, as evidenced by the following examples:

4) Recognition of the spoken word, for various speakers, (e.g. human voice to computer communication)

5) Recognition of a speaker regardless of the words spoken

6) Recognition of an environment or situation in which a system is placed. (Important for adaptive automatic control and learning systems)

7) Recognition of the location of faults in complex electronic systems

8) Character or signal transmission recognition over lines of communication, e.g. communication between computers

9) Target identification of aircraft, submarines, and missiles, and distinction from decoys using radar, sonar, etc.

10) Recognition of fields of agricultural crops, their condition, and state of growth from aerial observation.

The Pattern Recognition Device

The problem of designing devices which classify patterns requires two main investigations. The first investigation involves the problem often referred to as "feature extraction," i.e., operations on the pattern which determine its significant characteristics. The second investigation involves the decision-making device which classifies the pattern on the basis of the comparison of its characteristics (both similarities and differences) with those of a reference set of patterns. We now look at each of these problems in more detail.

Generally, in a pattern recognition problem a number of measureable

quantities exist which are used to characterize the patterns. The optimum choice of these quantities (called features) represents the "feature selection" problem mentioned above. This problem, by no means a trivial one, is as yet unsolved, and in fact is the major stumbling block to the total unification of all the applications of pattern recognition. Often the designer must use his intuition based on some prior experience to choose what seems to be a suitable set of features. On the basis of these features, studies are undertaken to determine the best decision or classifying strategies to employ.

In accordance with this subdivision of the pattern recognition problem into two subproblems, the recognition device is generally designed in two parts, one part being called the receptor and the second being the categorizer . A simple block diagram is shown in Figure 1.

The input to the receptor is the pattern to be recognized. The receptor, using various sensors, performs the task of measuring the chosen features. The output of the receptor is a vector (called the measurement of feature vector), whose components denote the various feature measurements.

The categorizer portion of the recognition device is responsible for assigning a given input pattern to a class, on the basis of the measurement vector. The designer constructs the categorizer to obtain the "best" possible recognition of the patterns to be classified. The term "best" used here refers to the best performance as indicated by the measure of goodness chosen by the designer. It should be noted that the optimum design of the categorizer in a particular problem is carried out with respect to a given set of features. To obtain the best overall system it is necessary to then optimize over all sets of possible features.

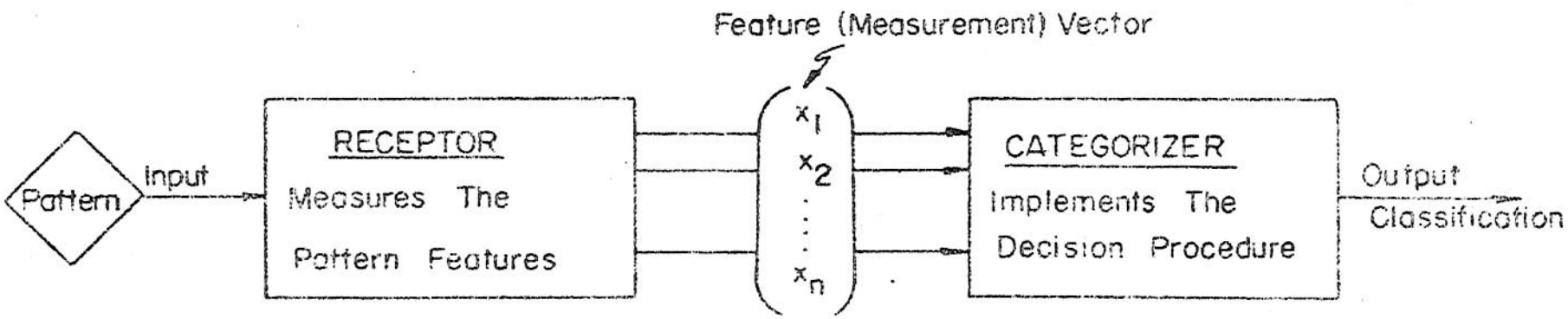


Figure 1
Block Diagram of a Pattern Recognition Device

An Example Problem

Perhaps the following example will be helpful in visualizing the operation of a pattern recognition device, and the function played by its components.

Consider the problem of remotely sensing whether a given field contains wheat, corn, or alfalfa. Assume we have decided that the percentage reflectance of electromagnetic energy in certain selected regions of the spectrum are the features. This choice of features could have been arrived at, for example, by examining the characterisitcs of the various crops as they would appear from the air. The receptor portion of the pattern recognition device then measures the percentage reflectance in the selected frequency bands.

Let x_1 be the percentage reflectance in band one, x_2 in band two .... x_n in band n where n is the total number of features measured.

The ordering of these features form a measurement vector $\underline{x} = (x_1, x_2 \ldots . x_n)$, and on the basis of this vector the categorizer is then to decide if the field is wheat, corn, or alfalfa.

We will examine this example further to introduce the concept of a measurement, or feature space, and then show how some of the common decision criteria can be represented in this space.

In order to represent a feature space easily on the plane of the paper, let us consider the situation in which we only measure two features (i.e., reflectance in two spectral bands). Thus, the feature vector contains only two components $\underline{x} = (x_1, x_2)$.

The receptor then represents each field examined (at its input) by two numbers (at its output). To start the process we might examine 10 fields each of wheat, corn, and alfalfa, then plot and label the classification of

each sample. Figure 2 shows a set of results which might be obtained. These known samples then constitute the reference set of patterns to which the patterns of unknown fields are compared. Obviously this reference set must be large enough and carefully selected, so that the set is typical of all future patterns to be classified. In practice the selection of this set is crucial, and requires great care and judgment. The point labeled U in Figure 2 represents the feature vector of an unknown field whose classification is to be determined.

The job of the categorizer begins at this point. That is, to classify the unknown field on the basis of its representation in the chosen feature space. Many methods for making the classification have been proposed and studied in the technical literature. We will mention only a few here to illustrate the approach.

1. Minimum distance to the means criterion - According to this approach the mean of each known class is found and represented as a point in the feature space. See Fig. 3. The pattern is then classified into the class whose mean is closest. This criterion then divides the space into three regions for classification as shown in Figure 3. Then, depending on whether the feature vector for the unknown field falls in regions A, C, or W, it is classified as alfalfa, corn, or wheat, respectively.

2. Minimum distance to the nearest member of a class - According to this criterion the distance from the unknown pattern to each reference pattern for each class is determined, and the minimum distance found. The unknown pattern is then classified into the same class as that of the reference pattern nearest it. As before, this decision criterion divides the feature space into decision regions. A graphical representation of this is shown in Figure 4.

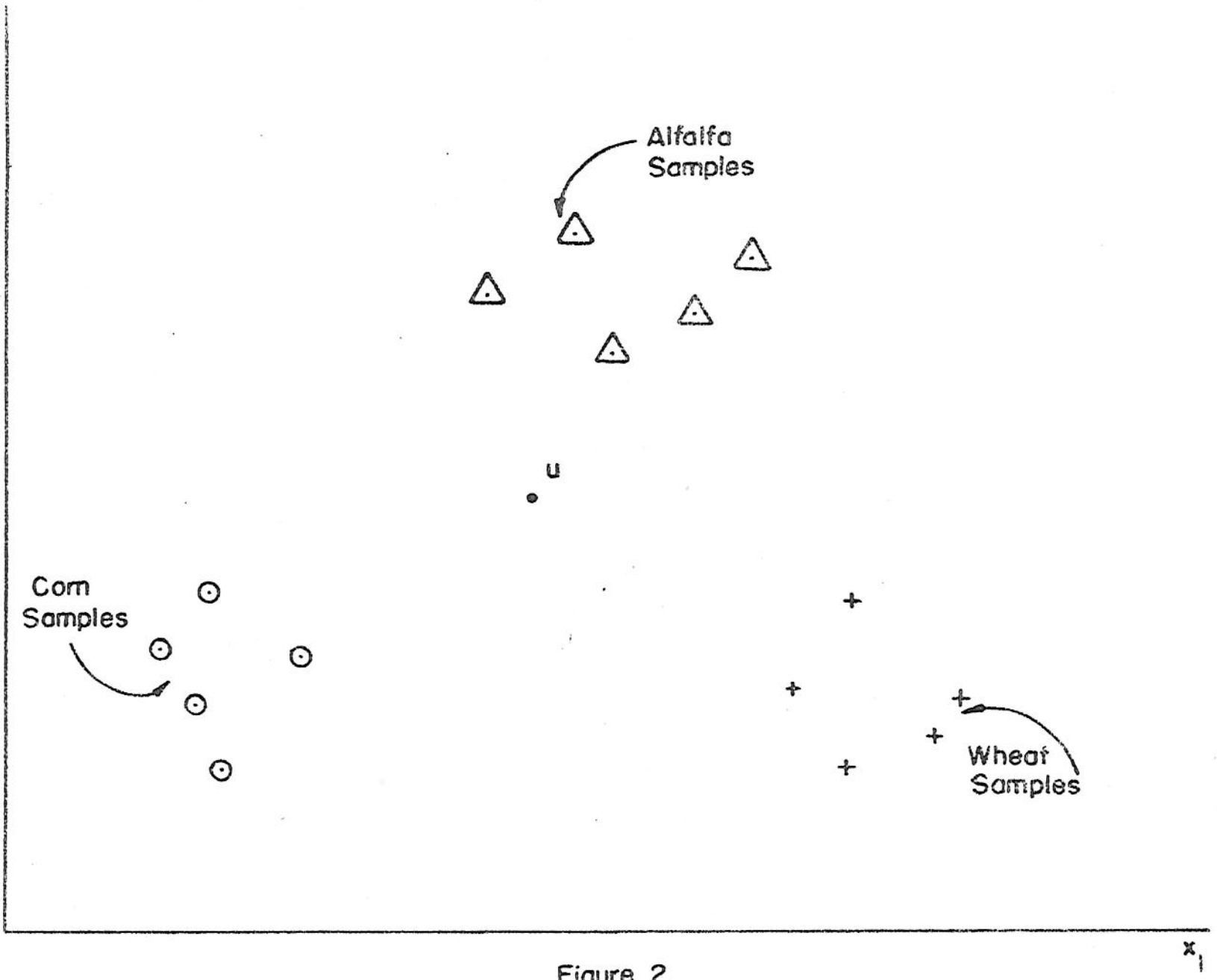


Figure 2

Samples in Two Dimensional Feature Space

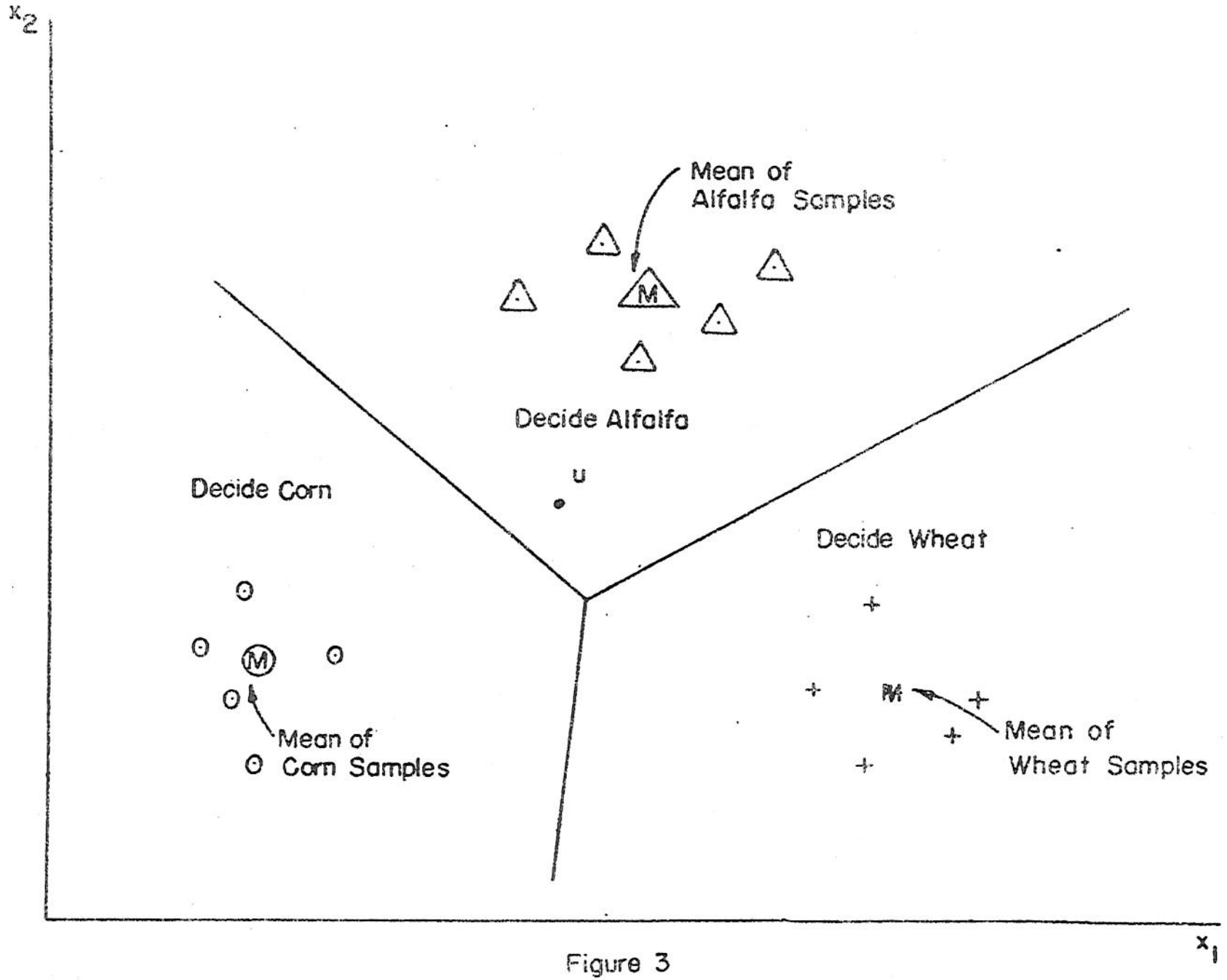


Figure 3

Decision Regions of Minimum Distance to Means Criterion

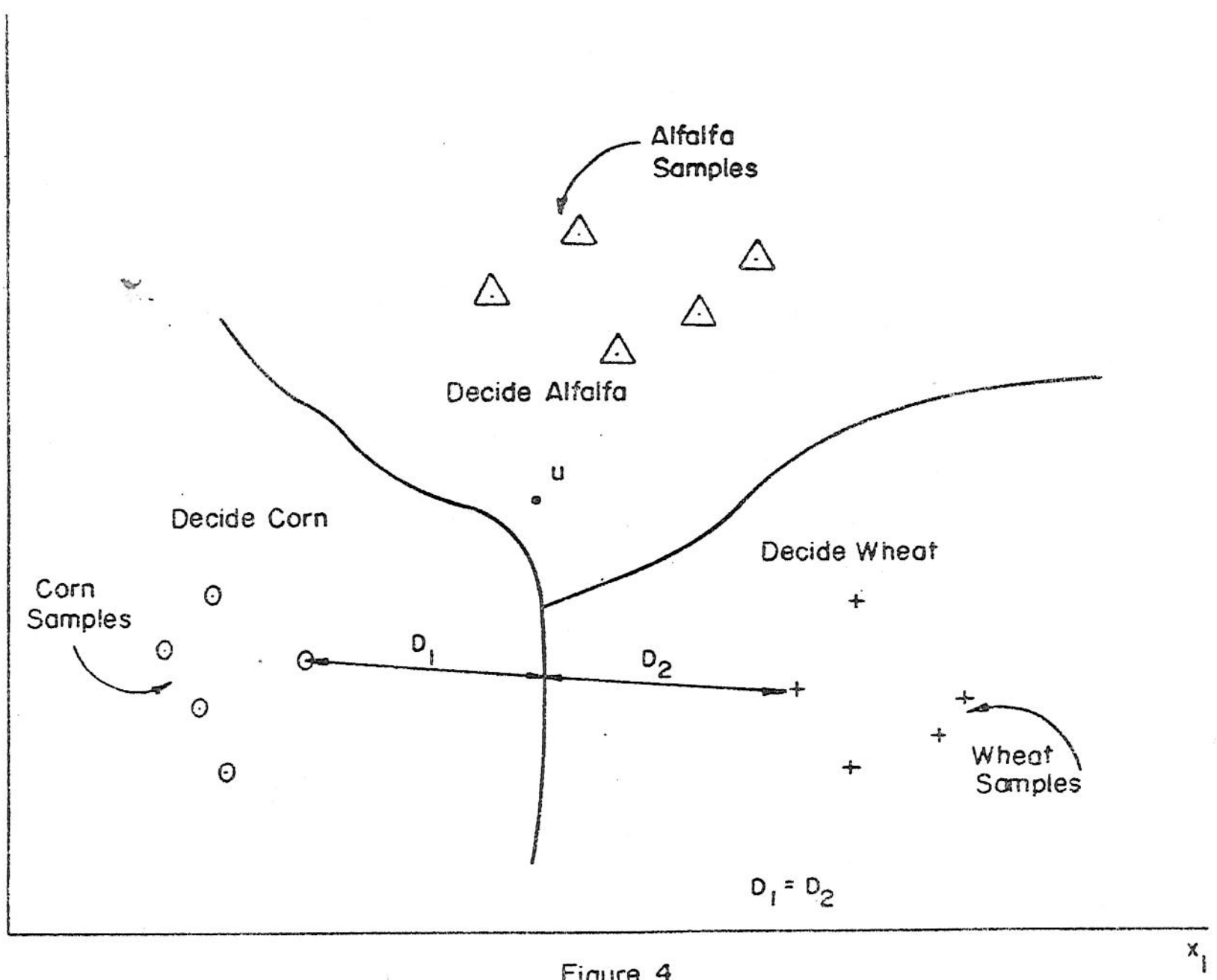


Figure 4

Decision Regions of Minimum Distance to Nearest Class Criterion

In these first two classification schemes, the assumption is made that the classes are sufficiently represented (characterized) by the limited number of known reference sample patterns. Specifically, in the example being considered it is assumed that the 10 reference samples of each class are sufficient to characterize the various classes. The justification of this assumption is a problem in its own right. As will be seen, this same assumption is employed in the third method of classification to be discussed, but in a slightly different way.

3. Statistical pattern recognition - Assume for the moment that we have three joint probability density functions, one each for the classes wheat, corn, alfalfa. Let each represent the probability that the representation in feature space of a sample of the particular class falls in a given region of the feature space. Therefore, we might have the three density functions shown in Figure 5. For each category of interest, a set of likelihood ratios can be computed. These likelihood ratios express the relative probabilities that a candidate point in question belongs to the category of interest rather than to any of the others. Thus, points in the feature space are assigned to the class for which the probability of occurrence of that point is the largest. Figure 6 shows the decision regions which might be obtained by this approach. It should be noted that most, if not all, optimum statistical decision criteria can be put in the form of a ratio criterion such as this.

We return for a moment to the problem of obtaining the joint distribution functions, and see how this is connected with the basic assumption discussed above. The knowledge of the density functions could come about in one of two ways: 1) The probability densities are actually known, say through some theoretical study. 2) If the probability densities are not known, this know-

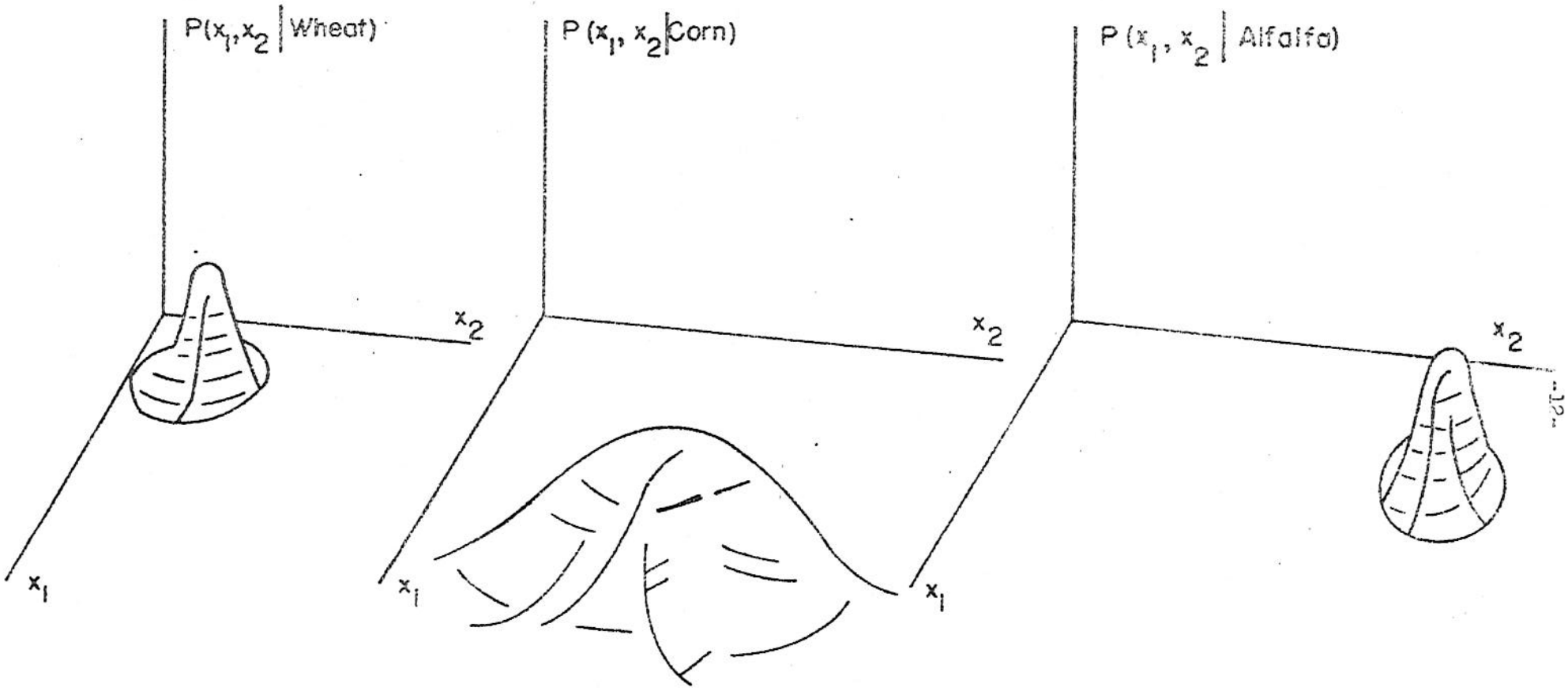
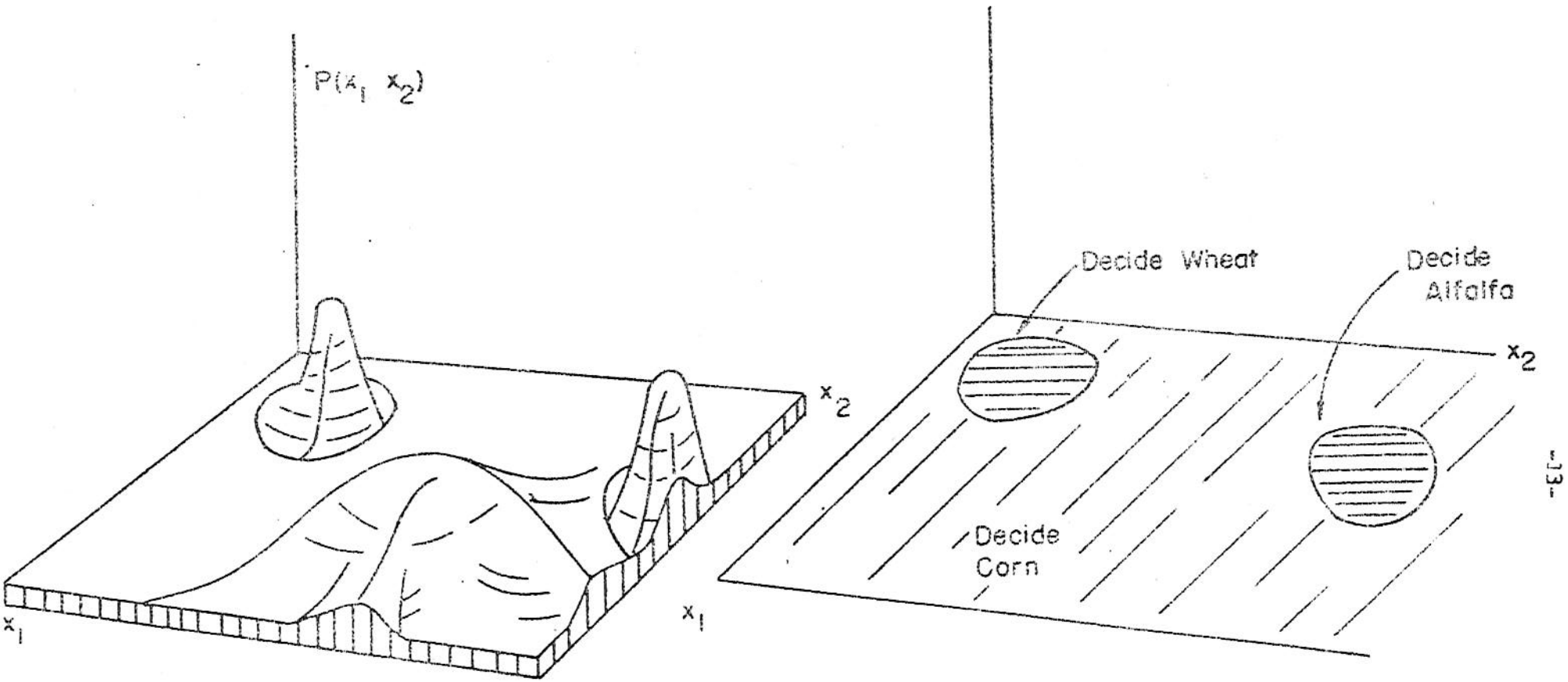


Figure 5
Probability Density Functions for Each Class



(a) Density Functions (b) Decision Regions for a Statistical Approach

Figure 6
Density Functions and Decision Regions for a Statistical Approach

ledge must be gained by taking samples of each class. Again a basic assumption which is employed is that the samples are sufficient to characterize the classes. In this case this means that the sample size is large enough to construct the probability densities required.

Using the geometrical interpretation of making classification decisions outlined above, we can summarize the discussion of the operation of the categorizer as follows. The feature space in which patterns are represented by points is divided into non-overlapping regions, one region corresponding to each of the categories. A classification decision consists of assigning to each candidate pattern the name of the category associated with the region in which the pattern is located.

Again in practice considerable judgment and experimentation are usually required to select the best categorizer approach for a given recognition task. A given categorizer may work well on one problem and not on another. Or two different ones may have the same error rates, but one may make different types (more costly) of errors and for different reasons.

A More Realistic Example

The above example was a hypothetical one designed to illustrate the approach. Let us now consider a more realistic situation. It is not usually possible in pattern recognition problems to conceive and design a receptor which is so effective that the various pattern classes are so obviously separable as they appear to be in Figure 2, at least not in only two dimensions.

Figure 7 illustrates a more typical situation, but again only in two dimensions in order to preserve the illustrative simplicity. The data for this illustration was obtained from actual measurements of the reflectance

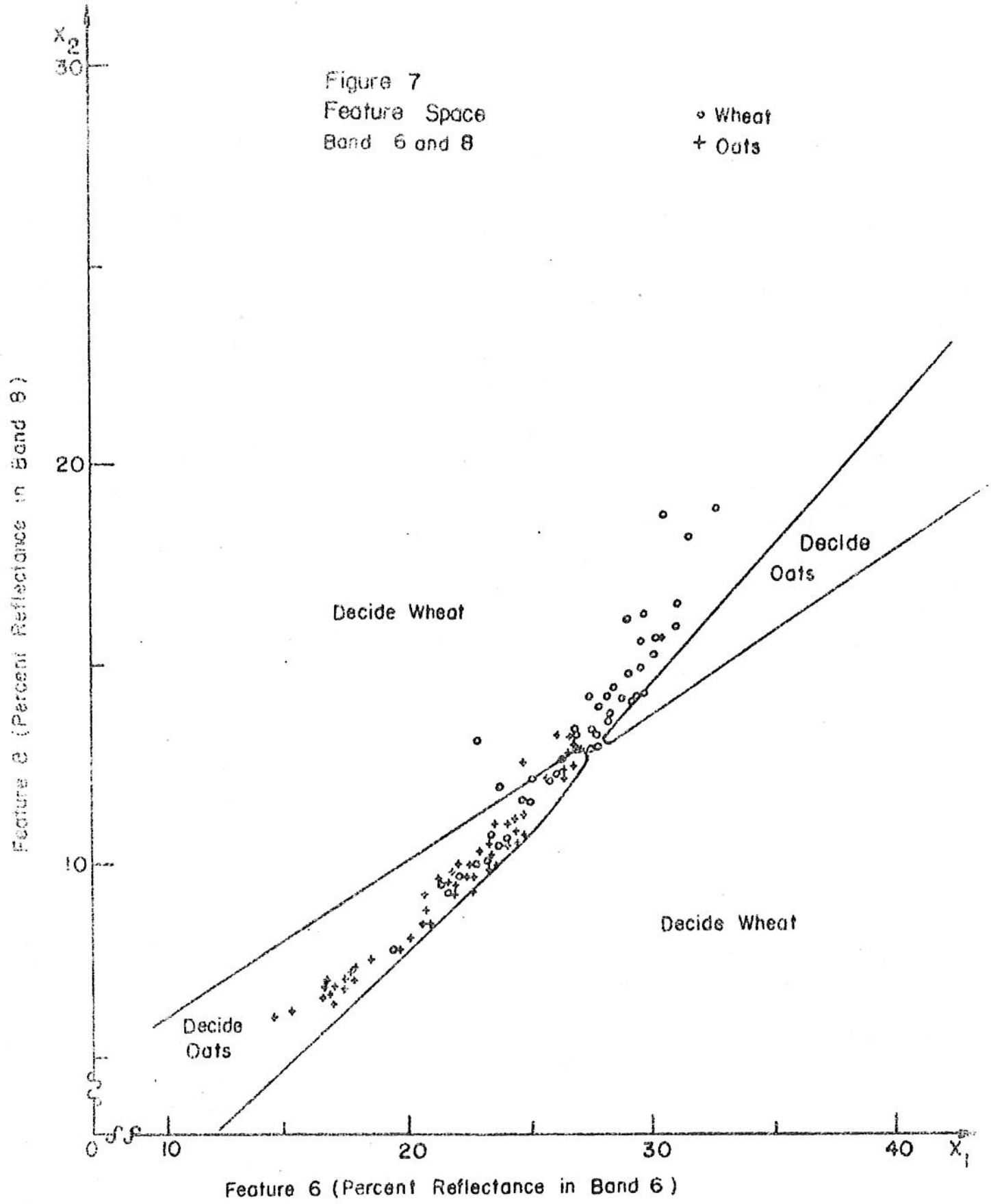


Figure 7
Feature Space
Band 6 and 8
∘ Wheat
+ Oats
Decide Wheat
Decide Oats
Decide Wheat
Decide Oats
Feature 8 (Percent Reflectance in Band 8)
Feature 6 (Percent Reflectance in Band 6)
X_2
X_1
0
10
20
30
40

of wheat and oats in two spectral bands. These two bands, numbers 6 and 8, were selected from nine for which data happened to have been available. Notice that the two pattern classes are somewhat overlapping in two dimensional space, indicating that perfect classification will not be possible.

An even worse situation is illustrated in Figure 8. The data plotted here is that from spectral bands 3 and 5 of the same wheat and oat reflectance measurements used in the previous figure.

In order to design a pattern recognition system, it is necessary to have a quantitative measure of the effectiveness of the receptor. For example, in the above case data is available from nine spectral bands. Suppose that we are limited to the use of only two of the nine bands. A limitation of this type could come about due to limits in computer speed or memory size*. The question then arises: what pair of features would be the best to use?

A few such measures of receptor effectiveness are already available in the technical literature. One, the divergence criteria, was used in the generation of this example. Divergence is a quantity defined for pairs of n-dimensional probability density functions. It yields a number which is proportional to the distance or separation between two densities; that is, the larger the divergence, the greater the separation.

In preparing this example, the divergence for each possible pair of the nine available features was computed. The feature pair with the highest divergence (bands 6 and 8) and a feature pair with a somewhat lower divergence (bands 3 and 5) were selected for presentation here.

---

*Actually, the size and speed of the average commercially available computer usually permits maximum dimensionalities of from 30 to several thousand, depending upon the particular categorizer algorithm.

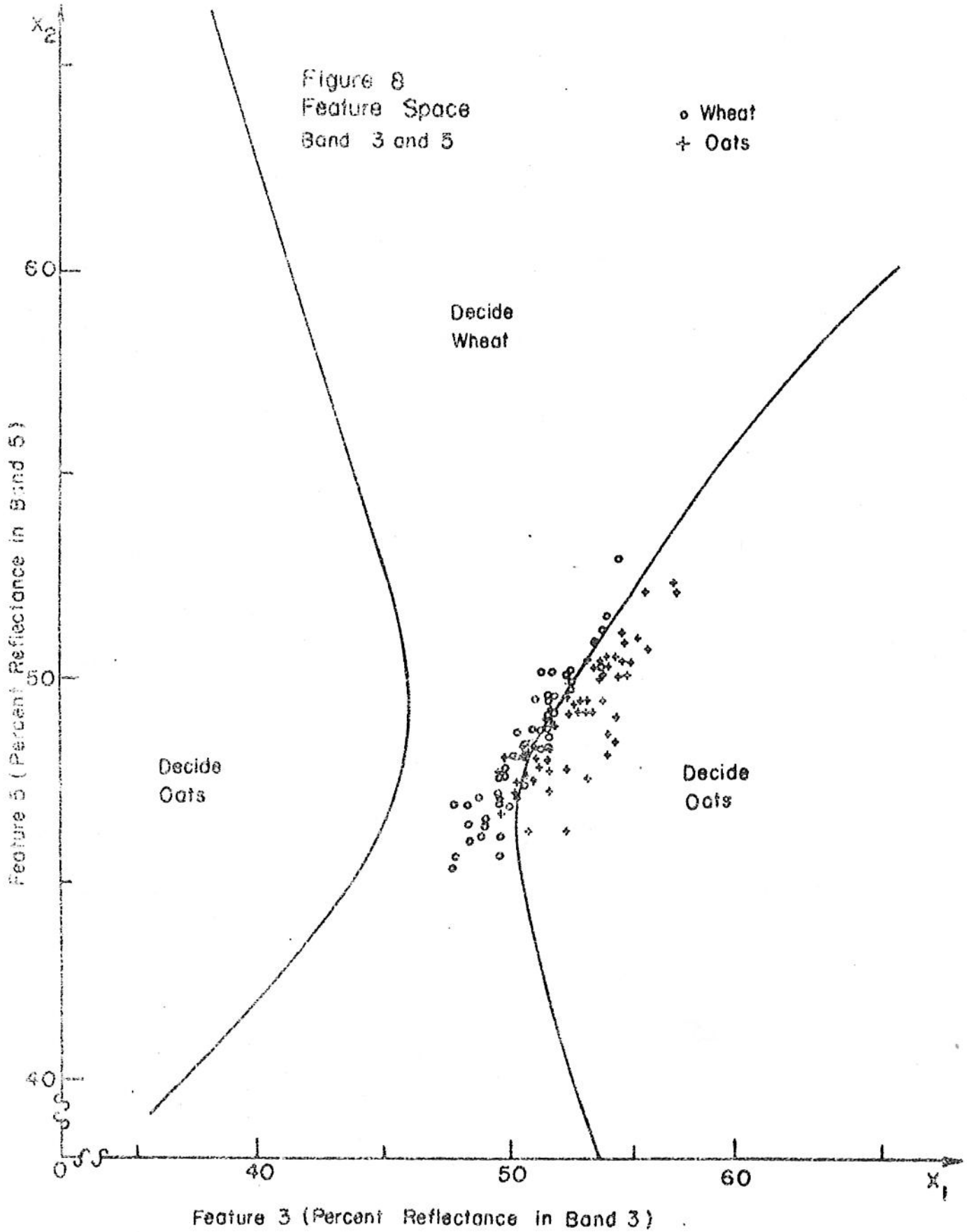


Figure 8
Feature Space
Band 3 and 5
o Wheat
+ Oats
Decide
Wheat
Decide
Oats
Decide
Oats
Feature 5 (Percent Reflectance in Band 5)
Feature 3 (Percent Reflectance in Band 3)
X_2
X_1
60
50
40
0
40
50
60

To complete this example, these samples of oats and wheat were classified using a maximum likelihood ratio categorizer (see Figure 6) with the above two feature pairs, and assuming Gaussian statistics. The results are shown in Tables I and II, and the decision boundaries (which turned out to be two-sheeted hyperbolas) are shown in Figure 7 and 8. To improve the classification performance over that of Table I, the designer could (a) further optimize the receptor in some fashion, (b) find a categorizer more suited to this specific problem, and/or (c) go to a higher dimensionality. Actually, all nine features could easily have been used.

TABLE I

Features 6 & 8

True Class	No. of Samples	No. classed as Oats	No. classed as Wheat	Percent Correct Classification
Oats	99	76	23	76.8
Wheat	78	11	67	85.9
			Overall	80.07

TABLE II

Features 3 & 5

True Class	No. of Samples	No. classed as Oats	No. classed as Wheat	Percent Correct Classification
Oats	99	73	26	73.7
Wheat	78	33	45	57.7
			Overall	66.6