MEMORANDUM
RM-4307-PR
OCTOBER 1964

ABSTRACTION AND PATTERN CLASSIFICATION

R. Bellman, R. Kalaba and L. A. Zadeh

PREPARED FOR:
UNITED STATES AIR FORCE PROJECT RAND

The RAND Corporation
SANTA MONICA • CALIFORNIA

MEMORANDUM
RM-4307-PR
OCTOBER 1964

ABSTRACTION AND PATTERN CLASSIFICATION

R. Bellman, R. Kalaba and L. A. Zadeh

This research is sponsored by the United States Air Force under Project RAND—Contract No. AF 49(638)-700 monitored by the Directorate of Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of the United States Air Force.

DDC AVAILABILITY NOTICE
Qualified requesters may obtain copies of this report from the Defense Documentation Center (DDC).

The RAND Corporation
1700 MAIN ST. • SANTA MONICA • CALIFORNIA • 90406

PREFACE

Part of the Project RAND research program consists of basic supporting studies in mathematics. In this Memorandum the authors formulate some mathematical concepts for dealing with the problem of pattern classification, which plays an important role in communication and control theories.

SUMMARY

This is a preliminary paper in which the authors discuss a general framework for the treatment of pattern-recognition problems. They make precise the notion of a "fuzzy" set. Then they show how this may be employed in a sequential experimental procedure to ascertain whether a symbol is a member of a particular set or not. The close relation between the problem of pattern recognition and interpolation is stressed.

CONTENTS

ABSTRACTION AND PATTERN CLASSIFICATION

1. INTRODUCTION

This Memorandum deals in a preliminary way with several concepts and ideas which have a bearing on the problem of pattern classification—a problem which plays an important role in communication and control theories.

There are two basic operations: abstraction and generalization, which appear under various guises in most of the schemes employed for classifying patterns into a finite number of categories. Although abstraction and generalization can be defined in terms of operations on sets of patterns, a more natural as well as more general framework for dealing with these concepts can be constructed around the notion of a "fuzzy" set—a notion which extends the concept of membership in a set to situations in which there are many, possibly a continuum of, grades of membership.

To be more specific, a fuzzy set A in a space $\Omega = \{x\}$ is represented by a characteristic function f which is defined on Ω and takes values in the interval [0,1], with the value of f at x, f(x), representing the "grade of membership" of x in A. Thus, if A is a set in the usual sense, f(x) is 1 or 0 according as x belongs or does not belong to A. When A is a fuzzy set, then the nearer the value of f(x) to 0, the more tenuous is the

membership of x in A, with the "degree of belonging" increasing with increase in f(x). In some cases it may be convenient to concretize the belonging of a point to a fuzzy set A by selecting two levels ε_1 and ε_2 ($\varepsilon_1, \varepsilon_2 \in [0,1]$) and agreeing that (a) a point x "belongs" to A if $f(x) \geq 1 - \varepsilon_1$; (b) x does not belong to A if $f(x) \leq \varepsilon_2$; and (c) x is indeterminate relative to A if $\varepsilon_2 < f(x) < 1 - \varepsilon_1$. In effect, this amounts to using a three-valued characteristic function, with $f(x) = 1$ if $x \in A$; $f(x) = 1/2$, say, if x is indeterminate relative to A; and $f(x) = 0$ if $x \notin A$.

Let A and B be two fuzzy sets in the sense defined above, with f_A and f_B denoting their respective characteristic functions. The union of A and B will be denoted in the usual way as

$$C = A \cup B, \tag{1}$$

with the characteristic function of C defined by

$$f_C(x) = \text{Max}(f_A(x), f_B(x)). \tag{2}$$

For brevity, the relation expressed by (2) will be written as

$$f_C = f_A \vee f_B. \tag{3}$$

Note that when A and B are sets, (2) reduces to the definition of "or."

In a similar fashion, the intersection of two fuzzy sets A and B will be denoted by

(4) $C = A \cap B$

with the characteristic function of C defined by

(5) $f_C(x) = \text{Min}(f_A(x), f_B(x))$,

which for brevity will be written as

(6) $f_C = f_A \wedge f_B$.

In the case of the intersection, when A and B are sets, (5) reduces to the definition of "and." When the characteristic functions are three-valued, (2) and (5) lead to the three-valued logic of Kleene [1].

2. ABSTRACTION AND GENERALIZATION

Let $x^1,\ldots,x^n$ be given members of a set A in Ω. In informal terms, by abstraction of $x^1,\ldots,x^n$ is meant the identification of those properties of $x^1,\ldots,x^n$ which they have in common and which, in aggregate, define the set A.

The notion of a fuzzy set provides a natural as well as convenient way of giving a more concrete meaning to the notion of abstraction. Specifically, let f^i denote the value of the characteristic function, f, of a fuzzy set A at a point x^i in Ω. A collection of pairs $\{(x^1,f^1),\ldots,(x^n,f^n)\}$ or, for short $\{(x^i,f^i)\}^n$, will be called a collection of samples or observations from A. By an abstraction on the collection $\{(x^i,f^i)\}^n$, we mean the estimation of the characteristic function of A from the samples $(x^1,f^1),\ldots,(x^n,f^n)$. Once an estimate of f

has been constructed, we perform a generalization on the collection $\{(x^i,f^i)\}^n$ when we use the estimate in question to compute the values of f at points other than $x^1,\ldots,x^n$.

An estimate of f employing the given samples $(x^1,f^1),\ldots,(x^n,f^n)$ will be denoted by $\tilde{f}$ or, more explicitly, by $(x;\{x^i,f^i)\}^n)$, and will be referred to as an abstracting function. Clearly, the problem of determining an abstracting function is essentially one of reconstructing a function from the knowledge of its values over a finite set of points. To make this problem meaningful, one must have some a priori information about the class of functions to which f belongs, such that this information in combination with the samples from A would be sufficient to enable one to construct a "good" estimate of f. As in interpolation theory, this approach involves choosing—usually on purely heuristic grounds—a class of estimates of f: $\tilde{F} = \{\tilde{f}(x\ ;\ \lambda) \mid \lambda \in R^e\}$ and finding that member of this family which fits, or fits "best" (in some specified sense of "best"), the given samples $(x^1,f^1),\ldots,(x^n,f^n)$. A special case of this procedure which applies to ordinary rather than fuzzy sets is the widely used technique for distinguishing between two sets of patterns via a separating hyperplane. Stated in terms of a single set of patterns, the problem in question is essentially that of finding, if it exists, a hyperplane L passing through the origin of $R^e(\Omega = R^e$, by assumption) such that the given points $x^1,\ldots,x^n$ belonging to a set A are all on

the same side of the hyperplane. (Note that, since A is a set, $f^1 = f^2 = \ldots = f^n = 1$.) In effect, in this case $\tilde{f}(x,\lambda)$ is of the form

$$(7) \quad \tilde{f}(x\ ;\ \lambda) = 1 \qquad \text{for } \langle x,\lambda\rangle \geq 0,$$

$$\tilde{f}(x\ ;\ \lambda) = 0 \qquad \text{for } \langle x,\lambda\rangle < 0,$$

where $\langle x,\lambda\rangle$ denotes the scalar product of x and λ, and the problem is to find a λ in R^e such that

$$(8) \quad \langle x^i,\lambda\rangle \geq 0 \qquad \text{for } i = 1,\ldots,n.$$

Any $\tilde{f}(x;\lambda)$ whose λ satisfies (8) will qualify as an abstracting function, and the corresponding generalization on $(x^1, 1),\ldots,(x^n, 1)$ will take the form of the statement "Any x satisfying $\langle x,\lambda\rangle \geq 0$ belongs to the same set as the samples $x^1,\ldots,x^n$." If one is not content with just satisfying (8) but wishes, in addition, to maximize the distance between L and the set of points $x^1,\ldots,x^n$ (in the sense of maximizing Min $\langle x^i,\lambda\rangle$), $||\lambda|| = 1$, then the determination of the corresponding abstracting function requires the solution of a quadratic program, as was shown by Rosen[2] in connection with a related problem in pattern recognition.

In most practical situations, the a priori information about the characteristic function of a fuzzy set is not sufficient to construct an estimate of f(x) which is "optimal" in a meaningful sense. Thus, in most instances

one is forced to resort to a heuristic rule for estimating f(x), with the only means of judging the "goodness" of the estimate yielded by such a rule lying in experimentation. In the sequel, we shall describe one such rule for pattern classification and show that a special case of it is equivalent to the "minimum-distance" principle which is frequently employed in signal discrimination and pattern recognition.

3. PATTERN CLASSIFICATION

For purposes of our discussion, a pattern is merely another name for a point in Ω, and a category of patterns is a (possibly fuzzy) set in Ω. When we speak of pattern classification, we have in mind a class of problems which can be subsumed under the following formulation and its variants.

Let A and B denote two* disjoint sets in Ω representing two categories of patterns. Suppose that we are given n points (patterns) $\alpha^1,\ldots,\alpha^n$ which are known to belong to A, and m points $\beta^1,\ldots,\beta^m$ which are known to belong to B. The problem is to construct estimates of the characteristic functions of A and B based on the knowledge of the samples $\alpha^1,\ldots,\alpha^n$ from A and $\beta^1,\ldots,\beta^m$ from B.

Clearly, one can attempt to estimate f_A without making any use of the β^j, $j = 1,\ldots,m$. However, in general,

*The restriction to two sets serves merely to simplify the analysis and does not entail any essential loss in generality.

such an estimate would not be as good as one employing both α's and β's. This is a consequence of an implied or explicit dependence between A and B (e.g., the disjointness of A and B), through which the knowledge of β's contributes some information about f_A. The same applies to the estimation of f_B.

The heuristic rule suggested in the sequel is merely a way of constructing estimates of f_A and f_B, given $\alpha^1,\ldots,\alpha^n$, and $\beta^1,\ldots,\beta^m$, in terms of estimates of f_A and f_B, given a single pair of samples α^i and β^j. Specifically, suppose that with every $\alpha \in A$ and every $\beta \in B$ are associated two sets $\tilde{A}(\alpha;\beta)$ and $\tilde{B}(\beta\ ;\ \alpha)$ representing the estimates of A and B, given α and β. (In effect, $\tilde{A}(\alpha\ ;\ \beta)$ defines the set of points in Ω over which the estimate $\tilde{f}_A(\alpha\ ;\ \beta)$ of f_A is unity, and likewise for $\tilde{f}_B(\beta\ ;\ \alpha)$ and $\tilde{B}(\beta\ ;\ \alpha$.) Points in Ω which are neither in $\tilde{A}(\alpha\ ;\ \beta)$ nor in $\tilde{B}(\beta\ ;\ \alpha)$ have indeterminate status relative to these sets.)

In terms of the sets in question, the estimates of A and B (or, equivalently, f_A and f_B), given $\alpha^1,\ldots,\alpha^n$ and $\beta^1,\ldots,\beta^m$, are constructed as follows

$$\tilde{A} = \bigcap_{j=1}^{m} \bigcup_{i=1}^{n} \tilde{A}(\alpha^i\ ;\ \beta^j), \tag{9}$$

$$\tilde{B} = \bigcap_{i=1}^{n} \bigcup_{j=1}^{m} \tilde{B}(\beta^j\ ;\ \alpha^i)\ . \tag{10}$$

Thus, under the rule expressed by (9) and (10), we generalize on $\alpha^1,\ldots,\alpha^n$ and $\beta^1,\ldots,\beta^m$ by identifying A with $\tilde{A}$ and

B with $\tilde{B}$. Note that this rule is consistent in the sense that if α is known to belong to A then $\alpha \in \tilde{A}$, and likewise for a point belonging to B. However, the consistency of this rule does not extend to fuzzy sets. Thus, if (9) and (10) were applied to the estimation of f_A and f_B when A and B are fuzzy sets, it would not necessarily be true that $f_A(\alpha) = \tilde{f}_A(\alpha)$ for all given α in A.

In essence, the rule expressed by (9) and (10) implies that a point x is classified as a member of A if and only if there exists an α^i such that for all β^j x lies in $\tilde{A}(\alpha^i,\beta^j)$. For this reason, the rule in question will be referred to as the "rule of complete dominance."

To illustrate the rule of complete dominance and indicate its connection with the "minimum-distance" principle which is frequently employed in signal discrimination, consider the simple case where Ω is R^e and $\tilde{A}(\alpha\ ;\ \beta)$ and $\tilde{B}(\beta\ ;\ \alpha)$ are defined as follows:

$$(11) \qquad \tilde{A}(\alpha\ ;\ \beta) = \left\{ x \,\middle|\, \prec x - \frac{(\alpha + \beta)}{2},\ \alpha - \beta > \succeq 0 \right\},$$

$$(12) \qquad \tilde{B}(\beta\ ;\ \alpha) = \left\{ x \,\middle|\, \prec x - \frac{(\alpha + \beta)}{2},\ \alpha - \beta > \prec 0 \right\} .$$

In effect, $\tilde{A}(\alpha\ ;\ \beta)$ is the set of all points which are nearer to α than to β or are equidistant from α and β, while $\tilde{B}(\beta\ ;\ \alpha)$ is the complement of this set with respect to R^e.

Now consider the following "minimum-distance" decision rule. Let A* and B* denote the sets of samples $\alpha^1,\ldots,\alpha^n$ and $\beta^1,\ldots,\beta^m$, respectively. Define the distance of a point x in Ω from A* to be $\underset{i}{\text{Min}} \; || \; x - \alpha^i \; ||$, where $|| \; ||$ denotes the Euclidean norm and $i = 1,\ldots,n$; do likewise for B*. Then, given a point x in Ω, decide that $x \in A$ if and only if the distance of x from A* is less than or equal to the distance of x from B*.

It is easy to show that this decision rule is a special case of (9) and (10). Specifically, with $\tilde{A}(\alpha;\ \beta)$ and $\tilde{B}(\beta\ ;\ \alpha)$ defined by (11) and (12), respectively, the decision rule in question can be expressed as follows:

$$(13)\quad x \in A \iff \forall_\beta{}^j \; \exists \alpha^i \left\{ || \; x - \alpha^i \; || \leq || \; x - \beta^j || \right\}, \quad i=1,\ldots,n, \quad j = 1,\ldots,m.$$

Now

$$(14)\qquad \tilde{A}(\alpha^i \; ; \; \beta^j) = \left\{ x \,\middle|\, || \; x - \alpha^i \; || \leq || \; x - \beta^j || \right\}$$

and consequently (13) defines the set

$$(15)\qquad \tilde{A} = \left\{ x \;\middle|\; \forall_\beta{}^j \; \exists \alpha^i (x \in \tilde{A}(\alpha^i \; ; \; \beta^j)) \right\} .$$

Clearly, (15) is equivalent to

$$(16)\qquad \tilde{A} = \bigcap_{j=1}^{m} \; \bigcup_{i=1}^{n} \; \tilde{A} \; (\alpha^i; \; \beta^j),$$

and similarly for B. Q.E.D.

In the foregoing discussion of the minimum-distance decision rule, we identified Ω with R^e and used the Euclidean metric in R^e to measure the distance between two patterns in Ω. However, in many cases of practical interest, Ω is a set of line patterns in R^2 such as letters, numerals, etc., to which the Euclidean metric is not applicable. In this case, the distance between two line patterns in R^2, say L_o and L_1, can be defined by

$$(17) \qquad d(L_o,L_1) = \underset{y_o \in L_o}{\text{Max}} \ \underset{y_1 \in L_1}{\text{Min}} \ || y_o - y_1 ||,$$

where $|| \ ||$ is the Euclidean norm in R^2, and y_o and y_1 are points in R^2 belonging to L_o and L_1, respectively.

Now suppose that we agree to regard two patterns L_o and L_1 as equivalent if one can be obtained from the other through translation, rotation, contraction (or dilation) or any combination of these operations. Thus, let T_δ denote the translation $y \to y + \delta$, where $y, \delta \in R^2$; let T_θ denote the rotation through an angle θ around the origin of R^2; and let T_ρ denote the contraction (or dilation) $x \to \rho x$ where $\rho \in R^1$. Then, we define the <u>reduced</u> distance of L_1 from L_o by the relation

$$(18) \quad d^*(L_1; L_o) = \underset{T_\delta}{\text{Min}} \ \underset{T_\theta}{\text{Min}} \ \underset{T_\rho}{\text{Min}} \ d(L_o, T_\delta T_\theta T_\rho L_1),$$

where $T_\delta T_\theta T_\rho L_1$ denotes the image of L_1 under the operation $T_\delta T_\theta T_\rho$ and $d(L_o, T_\delta T_\theta T_\rho L_1)$ is the distance between L_o and $T_\delta T_\theta T_\rho L_1$ in the sense of (17). Clearly, it is the reduced distance in the sense of (18) rather than the distance in the sense of (17) that should be used in applying the minimum-distance decision rule to the case where Ω is a set of line patterns in R^2.

To conclude our discussion of pattern classification, we shall indicate how the formulation given in the beginning of this section can be extended to fuzzy sets. Thus, let A and B denote two such sets in Ω, with f_A and f_B denoting their respective characteristic functions. Suppose that we are given n sample triplets $(x^1, f_A^1, f_B^1), \ldots, (x^n, f_A^n, f_B^n)$, with (x^i, f_A^i, f_B^i) representing a sample consisting of x^i and the values of f_A and f_B at x^i. The problem of pattern classification in this context is essentially that of estimating the characteristic functions f_A and f_B from the given collection of samples. Clearly, this formulation of the problem includes as a special case the pattern-classification problem stated earlier for the case where A and B are sets in Ω.

REFERENCES

1. Kleene, S. C., Introduction to Metamathematics, D. Van Nostrand Co., Inc., New York, 1952, p. 334.

2. Rosen, J. B., Pattern Recognition by Convex Programming, Stanford University, Technical Report No. 30, June 1963.