INFORMATION AND CONTROL 20, 301–312 (1972)

A Definition of a Nonprobabilistic Entropy in the Setting of Fuzzy Sets Theory

A. DE LUCA AND S. TERMINI

Laboratorio di Cibernetica del C.N.R., Naples, Italy

A functional defined on the class of generalized characteristic functions (fuzzy sets), called "entropy", is introduced using no probabilistic concepts in order to obtain a global measure of the *indefiniteness* connected with the situations described by fuzzy sets. This "entropy" may be regarded as a measure of a quantity of information which is not necessarily related to random experiments.

Some mathematical properties of this functional are analyzed and some considerations on its applicability to pattern analysis are made.

1. INTRODUCTION

The *fuzzy sets* theory was introduced by Zadeh (1965) in order to provide a scheme for handling a variety of problems in which a fundamental role is played by an indefiniteness arising more from a sort of *intrinsic ambiguity* than from a *statistical variation.*

Zadeh's scheme and its generalization by Goguen (1967) are intended to represent the counterpart of ordinary set theory in the field of not well-defined problems. One could then think of taking fuzzy sets as a basis for a generalization of such mathematical theories as probability, topology and so on, whose classical versions are founded on ordinary set theory. This program has been fruitfully undertaken by Zadeh (1968) and Chang (1968).

An algebraic analysis of the previous theory has been made by De Luca and Termini (1970) in order to better understand its relationships with classical set theory and, especially, obtain suggestions for the construction of mathematical calculi. Some implications of this approach in the analysis of complex systems have been considered by De Luca and Termini (1971).

In this note we propose the introduction of a "measure of the degree of fuzziness" or "entropy" of a generalized set. The meaning of this quantity is

Copyright © 1972 by Academic Press, Inc.
All rights of reproduction in any form reserved.

quite different from the one of classical entropy because no probabilistic[1] concept is needed in order to define it. This function gives a global measure of the "indefiniteness" of the situation of interest.

This function may also be regarded as an *average intrinsic information* which is received when one has to make a decision (as in pattern analysis) in order to classify ensembles of objects (patterns) described by means of fuzzy sets.

2. Entropy of a Fuzzy Set

Let us consider a set I and a lattice L; any map from I to L is called *L-fuzzy set* (Goguen, 1967). The name "fuzzy sets" given to these maps arises from the possibility of interpreting them, as done by Zadeh (1965), as a generalization of the characteristic functions of classical set theory. However, in the following, to avoid ambiguities the word *fuzzy sets* will refer preferably to maps instead of to abstract generalized sets endowed with membership functions.

Let us denote by $\mathscr{L}(I)$ the class of all maps from I to L. It is possible to induce a lattice structure to $\mathscr{L}(I)$ by the binary operations $\vee$ and $\wedge$ associating to any pair of elements f and g of $\mathscr{L}(I)$ the elements $f \vee g$ and $f \wedge g$ of $\mathscr{L}(I)$, defined point by point as

$$\begin{aligned} (f \vee g)(x) &\equiv \text{l.u.b.}\{f(x), g(x)\} \\ (f \wedge g)(x) &\equiv \text{g.l.b.}\{f(x), g(x)\}; \end{aligned} \tag{1}$$

l.u.b. and g.l.b. denote respectively the *least upper bound* and the *greatest lower bound* of $f(x)$ and $g(x)$ in the lattice L.

In this paper we will consider L as coinciding with the unit interval on the real line $L \equiv [0, 1]$; in this case (1) becomes

$$\begin{aligned} (f \vee g)(x) &= \max\{f(x), g(x)\}, \\ (f \wedge g)(x) &= \min\{f(x), g(x)\}. \end{aligned} \tag{2}$$

We try to introduce, for every element, or "fuzzy set", $f \in \mathscr{L}(I)$ a *measure* of the *degree of its "fuzziness"*. We require of this quantity, which we shall

[1] Here and in the following, by using the word "probability" we shall only refer to the *frequentistic* interpretation, without taking into account other interpretations such as the *logical* or *subjectivistic* (see, for instance, Carnap, 1967).

denote by $d(f)$, that it must depend only on the values assumed by f on I and satisfy at least the following properties:

P_1 $d(f)$ must be 0 if and only if f takes on I the values 0 or 1.

P_2 $d(f)$ must assume the maximum value if and only if f assumes always the value $\frac{1}{2}$.

P_3 $d(f)$ must be greater or equal to $d(f^*)$ where f^* is any "sharpened" version of f, that is any fuzzy set such that $f^*(x) \geqslant f(x)$ if $f(x) \geqslant \frac{1}{2}$ and $f^*(x) \leqslant f(x)$ if $f(x) \leqslant \frac{1}{2}$.

From now on let I be a finite set; this assumption and some others that we will make in the following, simplify the mathematical formalism but may be suitably weakened in future generalizations. We note, however, that the finiteness of I corresponds to a large class of actual situations.

We begin by introducing on $\mathscr{L}(I)$ the functional $H(f)$, formally similar to the Shannon entropy although quite different conceptually, whose range is the set of nonnegative real numbers and defined as

$$H(f) \equiv -K \sum_{i=1}^{N} f(x_i) \ln f(x_i) \tag{3}$$

where N is the number of elements of I and K is a positive constant.

We have the following:

Proposition 1. *$H(f)$ is a nonnegative valuation* (Birkhoff, 1967) *on the lattice* $\mathscr{L}(I)$, i.e.,

$$H(f \vee g) + H(f \wedge g) = H(f) + H(g) \qquad \text{for all } f, g \text{ of } \mathscr{L}(I) \tag{4}$$

In fact, from definition (3) and by (2) it follows that

$$\begin{aligned} H(f \vee g) &= -K \sum_{i=1}^{N} \max\{f(x_i), g(x_i)\} \ln \max\{f(x_i), g(x_i)\}, \\ H(f \wedge g) &= -K \sum_{i=1}^{N} \min\{f(x_i), g(x_i)\} \ln \min\{f(x_i), g(x_i)\}. \end{aligned} \tag{5}$$

Breaking up the sums in (5) into two parts, one extended over all x such that $f(x) \geqslant g(x)$ and the other over all x such that $f(x) < g(x)$, and summing up the right and left sides of (5), (4) is obtained.

Let us now give the following definitions:

DEFINITION 1. The power of a fuzzy set f is the quantity

$$F \equiv \sum_{i=1}^{N} f(x_i). \tag{6}$$

If f is a classical characteristic function, F reduces to the ordinary power of a (finite) set.

DEFINITION 2. If f and g are two fuzzy sets their direct product is the fuzzy set over $I^{(2)} = I \times I$ given by

$$(f \times g)(x, y) \equiv f(x) \cdot g(y)$$

where $(\cdot)$ denotes the ordinary product.

If f and g take on only the values 0 and 1, the previous definition reduces to the usual one of direct product of sets in terms of characteristic functions.

The functional (3) exhibits a sort of *additive property*; in fact, one has

$$\begin{aligned} H(f \times g) &= -K \sum_{i,j=1}^{N} f(x_i) \cdot g(y_j) \ln[f(x_i) \cdot g(y_j)] \\ &= G \cdot H(f) + F \cdot H(g), \end{aligned} \tag{7}$$

F and G being the powers of f and g, respectively.

If $F = G = 1$, then (7) becomes

$$H(f \times g) = H(f) + H(g). \tag{8}$$

For example, this happens with the function of the type $\varphi(x) \equiv f(x)/F$. In this case, (3) is formally identical to the *entropy* of the *finite scheme*

$$\begin{pmatrix} x_1 \cdots x_n \\ f(x_1) \cdots f(x_n) \end{pmatrix},$$

where $f(x_i)$ can be interpreted as the probability that x_i does occur in some random experiment.

One might be tempted to assume (3) as a measure of the fuzziness of a generalized set. We have then to see if $H(f)$ satisfies our requirements P_1, P_2 and P_3.

From definition (3) it follows that: $H(f) = 0$ if and only if f belongs to the subset of $\mathscr{L}(I)$ consisting of the classical characteristic functions (assuming, of course, $0 \cdot \ln 0 = 0$). Requirement P_1 is then satisfied. However, because the maximum of (3) is reached when $f(x) = 1/e$ for all x of I, in which case $H(f) = K \cdot N/e$, P_2 is not fulfilled.

It then seems more convenient to us to introduce the following functional, which we will call the "entropy" of the fuzzy set f:

$$d(f) \equiv H(f) + H(\bar{f}) \tag{9}$$

where $\bar{f}$, defined point by point as

$$\bar{f}(x) \equiv 1 - f(x)$$

satisfies the following noteworthy properties:

$$\begin{aligned} \bar{\bar{f}} &= f && \text{(involution law),} \\ \overline{f \vee g} &= \bar{f} \wedge \bar{g} && \\ & && \text{(De Morgan laws).} \\ \overline{f \wedge g} &= \bar{f} \vee \bar{g} && \end{aligned} \tag{10}$$

We explicitly note that $\bar{f}$, usually called the *complement* of f, is not the *algebraic complement* of f with respect to the lattice operations (2) (De Luca and Termini, 1970).

From (9) $d(f) = d(\bar{f})$; moreover, $d(f)$ can be written using Shannon's function $S(x) = -x \ln x - (1 - x) \ln(1 - x)$ as

$$d(f) = K \sum_{h=1}^{N} S(f(x_h)). \tag{11}$$

$d(f)$ satisfies requirements P_1 and P_2. Requirement P_3 is also satisfied. In fact, if f^* is a sharpened version of f we have by definition

$$(\alpha) \qquad 0 \leqslant f^*(x) \leqslant f(x) \leqslant \tfrac{1}{2}, \qquad \text{for} \quad 0 \leqslant f(x) \leqslant \tfrac{1}{2},$$

$$(\beta) \qquad 1 \geqslant f^*(x) \geqslant f(x) \geqslant \tfrac{1}{2}, \qquad \text{for} \quad \tfrac{1}{2} \leqslant f(x) \leqslant 1.$$

By the well-known property of Shannon's function $S(x)$—monotonically increasing in the interval [0, 1/2] and monotonically decreasing in [1/2, 1] with a maximum at $x = \frac{1}{2}$—we immediately get from (α) and (β) that, for any value of $f(x)$,

$$S(f^*(x)) \leqslant S(f(x)), \qquad x \in I.$$

From this relation by (11) it follows

$$d(f^*) \leqslant d(f).$$

We can now prove the

PROPOSITION 2. *$d(f)$ is a nonnegative valuation on the lattice $\mathscr{L}(I)$.*

In fact, from (9) and by (4) and (10) we have

$$\begin{aligned} d(f) + d(g) &= H(f \vee g) + H(f \wedge g) + H(\bar{f} \wedge \bar{g}) + H(\bar{f} \vee \bar{g}) \\ &= H(f \vee g) + H(\overline{f \vee g}) + H(f \wedge g) + H(\overline{f \wedge g}) \\ &= d(f \vee g) + d(f \wedge g). \end{aligned}$$

If we assume in (11) that $K = 1/N$, we obtain the functional

$$\nu(f) = \frac{1}{N} \sum_{h=1}^{N} S(f(x_h)) \tag{12}$$

which we will call the "normalized entropy". This name is appropriate because, taking the logarithm in base 2, one has

$$0 \leqslant \nu(f) \leqslant 1 \qquad \text{for all} \quad f \in \mathscr{L}(I).$$

By Proposition 2 it follows immediately that also $\nu(f)$ is a nonnegative valuation on the lattice $\mathscr{L}(I)$.

3. INTERPRETATION OF $d(f)$

In this section we will further discuss the meaning of the entropy $d(f)$ previously introduced, interpreting it also as "quantity of information."

The analysis of a particular example will help to understand better the next general definitions. Finally, the case in which the "incertitude" arises both from the "fuzziness" of the description and from statistical variations will be considered.

In the previous section, the functional $d(f)$ has been assumed as giving a *measure* of the fuzziness of f; this quantity, as we will see, may also be considered as measuring an *amount of information* even if its meaning is different from the standard one of Shannon's information theory.

Let us now discuss the following example. We consider N cells x_i $(i = 1 \cdots N)$ of sensory units (as photoelectric cells) disposed in a two-dimensional array I, or *retina*, and suppose, first, that one may project on the retina only patterns such that any cell can "see" only *white* or *black* colors, to which correspond two different states 1 or 0, of the photoelectric unit. Therefore, we may associate with the x_i cell $(i = 1,..., N)$ a variable $f(x_i)$

which may assume only two values: 1 (white) and 0 (black). In this way, to any pattern corresponds a subset of I, that is the one formed by the cells such that $f(x_i) = 1$.

Let us now suppose that for some reasons (generally depending on the projected pattern) whose nature at the moment does not interest us, the state $f(x_i)$ of the i-th cell can vary in the interval [0, 1] instead of the set {0, 1}; this means that the cell may "see" a discrete or continuous number of *grey* colors with each of which, by means of the photoelectric units, we associate a number that we interpret, according to the scale, as "degree of white" or "degree of black". In such a way a description of a pattern projected on the retina may be made by means of a fuzzy set f. This pattern, described by f, looks "ambiguous" to any "people" or "device" which knows only black or white; a measure of this ambiguity is $d(f)$. The nature of this ambiguity therefore arises from the "incertitude" present when we must decide, looking at the grey color of the i-th cell, if this has to be considered white or black. We may measure this incertitude by $S(f(x_i))$, which is 0 if $f(x_i)$ is equal to 0 or 1 and is maximum for $f(x_i) = 1/2$; the *total amount of incertitude* is $\sum_{i=1}^{N} S(f(x_i)) = d(f)$.

Now if we carry out some experiments by which we can remove or reduce the uncertainty which existed before the experiment, we can say that we obtain some *information*. Let us, first, assume that the experiment consists in taking a decision about the colors (white or black) of *all* the N cells of the retina. In this way we produce a new classical pattern f^*. This kind of experiment makes the ambiguity of the final pattern f^* equal to 0 by completely removing the *uncertainty* on the colors of the cells which existed before the experiment. It seems natural to us to assume that in these experiments we receive an average amount of information proportional or equal to (choosing some unit) the initial uncertainty $d(f)$.

We may assume that $d(f)$ also measures the average amount of information (about the colors of the pattern) which is lost going from a classical pattern to the fuzzy pattern f.

It is also possible to consider only *partial removals* of uncertainty in any experiment by which we transform the fuzzy pattern f into a new fuzzy pattern $\tilde{f}$ having

$$d(\tilde{f}) \leqslant d(f).$$

In this way we may say that we receive a quantity of information measured by $d(f) - d(\tilde{f})$.

We emphasize that the ambiguity we have previously defined and the related information are "structural", that is, linked to the fuzzy description, while in the classical information theory it is due to the uncertainty in the previsions of the results of random experiments.

Let us now consider any experiment in which the elements $x_1, \ldots, x_N$ of I may occur, one and only one in each trial, with probabilities $p_1, p_2, \ldots, p_N$ ($p_i \geqslant 0, \sum_{i=1}^{N} p_i = 1$). If a fuzzy set f is defined in I, we have two kinds of uncertainty:

(i) The first uncertainty of "random" nature is related to the prevision of the result, i.e., the element of I which will occur. As is well known, the average uncertainty is measurable by Shannon's entropy

$$H(p_1 \cdots p_N) = -\sum_{i=1}^{N} p_i \ln p_i;$$

H also gives the average information which is received knowing the element which occurs.

(ii) The second uncertainty of "fuzzy" nature concerns the interpretation of the result as 1 or 0. If the result is x_i we still have an amount of incertitude measurable by

$$S(f(x_i)).$$

The statistical average of $S(f(x_i))$

$$m(f, p_1 \cdots p_N) = \sum_{i=1}^{N} p_i S(f(x_i)), \tag{13}$$

which coincides with the normalized entropy (12) if $p_1 = p_2 = \cdots = p_N = 1/N$, represents the (statistical) average information received taking a decision (1 or 0) on the elements x_i ($i = 1 \cdots N$). This is an interesting new concept because it may happen, for instance, that the elements of I with $f(x) \approx \frac{1}{2}$ may occur in the random experiment only exceptionally where the elements with $f(x)$ near the bounds 1 or 0 may occur very frequently. In such a case m is small even if f is quite "soft"; this happens because the statistical uncertainty on the decisions is, in fact, small.

We can consider the total entropy

$$H_{\text{tot}} = H(p_1 \cdots p_N) + m(f, p_1 \cdots p_N), \tag{14}$$

which may be interpreted as the total average uncertainty that we have in *making a prevision* about the element of I which will occur as a result of the random experiment and in *taking a decision* about the value 1 or 0 which has to be attached to the element itself. If $m = 0$, which occurs in the absence of fuzziness, H_{tot} reduces itself to the classical entropy $H(p_1 \cdots p_N)$. If we have $H(p_1 \cdots p_N) = 0$, which means there is no random experiment and only a fixed element, say x_i, will occur, $H_{\text{tot}} = S(f(x_i))$.

We observe that the previous formula (14) is formally identical to one of the ordinary information theory

$$H(AB) = H(A) + H_A(B),$$

giving the entropy of the product scheme AB in the case in which the events of B are statistically dependent on those of A.

Another case which we may consider is when the fuzzy set f is *random*: that is, f is a map

$$f : \Omega \times I \to [0, 1]$$

such that, for any fixed x, $f(\xi, x)$ is a *random variable* with respect to a given probability space $(\Omega, \mathscr{F}, p)$ where Ω is the nonempty set of *sample points*, $\mathscr{F}$ a σ-field of subsets of Ω and p a *probability measure*. For any fixed ξ, $f(\xi, x)$ is a *fuzzy set*. Let us consider the case when Ω has only a finite number M of elements $\xi_1, \ldots, \xi_M$ which may occur with probabilities $p(\xi_1), \ldots, p(\xi_M)$; we may introduce an average fuzzy set $\langle f \rangle$ as

$$\langle f \rangle(x) \equiv \sum_{i=1}^{M} f(x, \xi_i)\, p(\xi_i).$$

In such a case the entropy of the fuzzy set is itself a random variable; in fact, if the event ξ_i happens, we have the fuzzy set $f(\xi_i, x)$ whose entropy is $d_i(f)$. In this case it is meaningful to consider the average entropy given by

$$\sum_{i=1}^{M} p(\xi_i)\, d_i = \sum_{i=1}^{M} \sum_{j=1}^{N} p(\xi_i)\, S(f(\xi_i, x_j)) \tag{15}$$

which reduces itself to (11) in the deterministic case.

4. Equivalence Classes in $\mathscr{L}(I)$

Let us now introduce, for any pair of elements f and g of $\mathscr{L}(I)$, the quantity

$$\delta(f, g) \equiv | d(f) - d(g) |$$

which satisfies the properties

$$\left.\begin{array}{r}\delta(f,f)=0\\ \delta(f,g)\geqslant 0;\qquad \delta(f,g)=\delta(g,f)\\ \delta(f,g)\leqslant\delta(f,h)+\delta(h,g)\end{array}\right\}\quad\text{for all } f,\ h,\ g \text{ of } \mathscr{L}(I).$$

$\mathscr{L}(I)$ is a *pseudo-metric* space with respect to δ, but not a metric one because $\delta(f,g)=0$ does not necessarily imply $f=g$. However, we may introduce on $\mathscr{L}(I)$ the equivalence relation

$$f\sim g \qquad \text{iff} \quad \delta(f,g)=0, \qquad \text{that is,} \quad d(f)=d(g),$$

and decompose $\mathscr{L}(I)$ into equivalence classes C_i, which means considering the quotient space $\mathscr{L}(I)/\sim$.

$\mathscr{L}(I)/\sim$ is a strict *metric space* with respect to the distance

$$\delta(C_i, C_j)\equiv|\,d(C_i)-d(C_j)|,$$

where $d(C_i)=d(f)$ with $f\in C_i$, and may be ordered in a chain, with respect to the relation

$$C_i\leqslant C_j \qquad \text{if and only if} \quad d(C_i)\leqslant d(C_j),$$

with a least element C_0 that is, the class of all classical characteristic functions ($d(C_0)=0$), and a greatest element C_1, that is, the class to which only the function which is always $1/2$ belongs; in this case, assuming $K=1$ and the logarithm in base 2, $d(C_1)=N$.

Finally, we note that if we consider an experiment in which it is possible to choose some fuzzy sets only from a finite number of the previous classes $C_1, C_2,\ldots, C_M$ with probabilities $p(C_1), p(C_2),\ldots, p(C_M)$ ($\sum_{i=1}^{M} p(C_i)=1$), the entropy of the chosen fuzzy set is, as in (15), a random variable which may assume the values $d(C_1),\ldots, d(C_M)$ with probabilities $p(C_1),\ldots, p(C_M)$. In this case it is natural to introduce the average entropy

$$\langle d\rangle=\sum_{i=1}^{M} p(C_i)\, d(C_i). \tag{16}$$

This situation may happen in pattern analysis when we consider the case of patterns, described by means of fuzzy sets, belonging only to a finite number of classes of entropies $d(C_1),\ldots, d(C_M)$, and occurring with probabilities $p(C_1),\ldots, p(C_M)$; this means that the previous patterns may have different levels of fuzziness (or ambiguity), a global average measure of which is given by (16).

5. Concluding Remarks

The considerations exposed in the previous sections are not complete and many mathematical generalizations are possible. As we said in Section 2, we have considered a universal class I with a *finite* number of elements; it would be interesting to extend the previous concepts for the infinite case. Furthermore, we have defined the entropy $d(f)$ only in the case in which the range of the generalized characteristic functions is the interval $[0, 1]$ of the real line which is a chain, i.e., all its elements are comparable. A possible extension of the concept of entropy of fuzzy sets, on which we are actually working, may be made in the case of L-fuzzy sets, where L is a lattice with universal bounds 0 and 1. We observe that the possibility of a general lattice as range of generalized characteristic functions is not hypothetical but naturally arises every time we consider propositions whose truth-values are not always comparable. It seems to us that such situations occur in pattern recognition when, in order to classify an object, one considers different properties of it which are not comparable as, for instance, weight and color.

The discussion made in Section 3 about the entropy $d(f)$ of a fuzzy set f shows that the decision theory plays, in this context, a rôle similar to the one of probability in information theory. However, in our case it would be desirable to prove some general theorems, as Shannon's ones of information theory, by which one may evaluate the actual implications of the given concept of entropy. We think that these theorems must relate some important aspects of taking decisions with the entropy of the described situations.

In conclusion, we wish to stress that, although many mathematical and interpretative problems are open, the previous concept of entropy, different from the classical one, may be regarded as a first step in the attempt to found a useful calculus in the context of fuzzy sets theory.

Acknowledgments

We acknowledge our debt to Professor E. R. Caianiello for many interesting discussions and to the referee for his useful suggestions.

Received: July 29, 1970

References

Birkhoff, G. (1967), "Lattice Theory," 3rd ed., Colloquium Publications, Vol. 25, American Mathematical Society, Providence, RI.

Carnap, R. (1967), "The Logical Foundations of Probability," 2nd ed., University of Chicago Press, Chicago, IL.

CHANG, C. L. (1968), Fuzzy topological spaces, *J. Math. Anal. Appl.* **24**, 182–190.

DE LUCA, A. AND TERMINI, S. (1970), Algebraic properties of fuzzy sets, abstract, *Notices Amer. Math. Soc.* **17**, 944; *J. Math. Anal. Appl.*, to appear.

DE LUCA, A. AND TERMINI, S. (1971), Algorithmic aspects in complex systems analysis, *Scientia* **106**, 659–671.

GOGUEN, J. A. (1967), *L*-fuzzy sets, *J. Math. Anal. Appl.* **18**, 145–174.

ZADEH, L. A. (1965), Fuzzy sets, *Information and Control* **8**, 338–353.

ZADEH, L. A. (1968), Probability measures of fuzzy events, *J. Math. Anal. Appl.* **23**, 421–427.