Observational entropy: entropy in the context of indistinguishability operators

Enric Hernàndez
Secció de Matemàtiques i Informàtica
ETSAB. Avda. Diagonal 649, 08028 Barcelona
Universitat Politècnica de Catalunya
enriche@lsi.upc.es

Jordi Recasens
Secció de Matemàtiques i Informàtica
ETSAV. Pere Serra 1-15, 08190 Sant Cugat
Universitat Politècnica de Catalunya
recasens@ea.upc.es

Summary

This work is intended to study the notion of entropy in the context of indistinguishability operators. The adapted notion of entropy should combine two kinds of uncertainty: the one coming from random (probabilistic uncertainty) and the uncertainty in the observation of the elements.

Keywords: entropy, observational entropy, Shannon's entropy, observation degrees, T-indistinguishability operators.

1 Introduction

The problem of measuring the uncertainty of a set of events is not new. The first attempts tried to quantify the uncertainty associated to a random experiment. So, Hartley (1928) captured the intuitive idea that the more possible results for an experiment, the less it can be predicted, defining the uncertainty as $\log(n)$, where n is the number of possible results. Then we have $\log(n) < \log(n+1)$, and also that the uncertainty associated with the product of two experiments is the sum of the single uncertainties ($\log(pq) = \log(p) + \log(q)$).Anyway, this measure had the drawback of ignoring the probability of the events. This difficulty was overcome by Shannon (1948) defining the entropy of a random variable as:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \quad (1)$$

It is important to note that this measure was thought within the frame of communication theory, specifically for facing issues concerning channel reliability and reduction of transmission cost, but ignoring the semantic content of the messages involved. What happens when events "carry" a concrete meaning, defined in terms of risk, utility, ..? [2] presents a situation worthwhile enough to be reproduced, in order to make clear that point: let us suppose two roads R_1,R_2 with probabilities of accident 0.28 and 0.72 respectively. Shannon's entropy is the same for both of them, but it seems quite obvious which one we would choose, based on the uncertainty of the risk of an accident. Therefore, providing the set of events with a particular semantics requires a "further step", in the sense that we need to adapt Shannon's measure in order to express random uncertainty in terms of this semantics. In this work, this semantics will be established by defining an indistinguishability relation among the elements of some domain, making some elements indistinguishable from others. The main idea is that the occurrence of two different events, but indistinguishable by the indistinguishability relation defined, will count as the occurrence of the same event when measuring the "observational" entropy.

Let us recall the definition of T-indistinguishability operator [3]:

Definition 1.1: A fuzzy relation E defined over X is a T-indistinguishability operator (being T a t-norm) if $\forall x, y, z \in X$:

- $E(x, x) = 1$ (reflexivity)
- $E(x, y) = E(y, x)$ (symmetry)
- $T(E(x, y), E(y, z)) \leq E(x, z)$ (T-transitivity)

2 Previous work

There has been done some work intended to combine random and fuzzy uncertainty. The proposals can be grouped in three main tendencies. In the sequel $P(X)$ will denote a probability distribution on X and A a fuzzy set on X with membership function μ_A.

- "Global" combination: algebraic combination (sum, product, ..) of the two kinds of uncertainty which, in turn, are calculated isolatedly. An exemple is the measure of Xie and Bedrosian (1984) [4]:

$$H(A,P) = -\sum_{x\in X} p(x)\log_2 p(x) - \sum_{x\in X}(\mu_A(x)\log_2\mu_A(x) + (1-\mu_A(x))\log_2(1-\mu_A(x))) \quad (2)$$

, where the first term measures the random entropy (Shannon's entropy), and the rest measures the fuzziness degree of A (by using DeLuca and Termini's (1972) expression).

- "Local" combination (element by element): each element combines its contribution to both kinds of uncertainty locally. The final value is the sum of all those individual contributions. An exemple is the DeLuca and Termini's measure:

$$H(A,P) = -\sum_{x\in X} p(x)\left[\mu_A(x)\log_2\mu_A(x) + (1-\mu_A(x))\log_2(1-\mu_A(x))\right] \quad (3)$$

- Weighted measures: one of the two values (probability or possibility degree) "weights" the contribution of the other. Zadeh's measure [6] exemplifies this trend:

$$H(A,P) = -\sum_{x\in X}\mu_A(x)p(x)\log_2 p(x) \quad (4)$$

,or the one found in [2]:

$$H(A,P) = -\sum_{x\in X}\frac{\mu_A(x)}{\sum_{x\in X}p(x)\mu_A(x)}p(x)\log_2 p(x) \quad (5)$$

3 Observational entropy

In [5] a new measure of entropy, suitable to operate on domains over which a similarity relation has been defined, is introduced. Our point of view is that this measure overcomes some limitations of the previous proposals. In this section we will recall some basic definitions and properties of this entropy measure, and will provide a new interpretation in terms of the observability of the elements.

Definition 3.1: Let E a T-indistinguishability operator on a set X. The observation degree of $x_j \in X$ is defined by:

$$\pi(x_j) = \sum_{x\in X} p(x)E(x,x_j). \quad (6)$$

Due to the reflexivity of E, this expression can be rewritten as:

$$\pi(x_j) = p(x_j) + \sum_{x\in X|x\neq x_j} p(x)E(x,x_j). \quad (7)$$

This definition has a clear interpretation: the possibility of observing x_j is given by the probability that x_j really happens (expressed by the first term), plus the probability of ocurrence of some element "very close" to x_j, weighted by the similarity degree . In other words, the first term measures the possibility of really observing x_j, while the second term measures the possibility of observing x_j mistakenly (x_j didn't really happen). It holds:

Lemma 3.2:

$$\forall x \in X : 0 \leq \pi(x) \leq 1. \quad (8)$$

Corollary 3.3:

$$0 \leq \sum_{x\in X}\pi(x) \leq Card(X). \quad (9)$$

It should be noted that $\pi(X)$ is not a probability distribution since $\sum_{x\in X}\pi(x) \neq 1$.

Definition 3.4: The quantity of information received by observing x_j is defined by:

$$C(x_j) = -\log_2\pi(x_j). \quad (10)$$

Definition 3.5: Given a T-indistinguishability operator E on X, and P a probability distribution on X, the observational entropy (HO) of the pair (E,P) is defined by:

$$HO(E,P) = \sum_{x\in X} p(x)C(x). \quad (11)$$

Let's suppose the following case: $X = \{x_1, x_2\}$, P the following probability distribution $p(x_1) = p(x_2) = 0.5$, and E the T-indistinguishability operator defined by $E(x_1,x_1) = E(x_2,x_2) = 1$; $E(x_1,x_2) = E(x_2,x_1) =$

0. It is trivial to check that $HO(E,P)=1$. This result suggests the following definition:

Definition 3.6: The information received by observing an event between two equally probable and fully distinguishable events will define the unit of measure for the observational entropy: the observable bit.

Lemma 3.7: Let E be a T-indistinguishability operator on X, P a probability distribution on X and $H(P)$ the Shannon's entropy.

$$HO(E,P) \leq H(P). \tag{12}$$

Lemma 3.8: Let E, E' be two T-indistinguishability operators on X and P a probability distribution on X.

$$E \leq E' \rightarrow HO(E,P) \leq HO(E',P). \tag{13}$$

Lemma 3.9: Given $X = \{x_1, x_2\}$, E the T-indistinguishability operator on X such that

$$E(x_i, x_j) = \begin{cases} 1 & , i = j \\ 0 & , i \neq j \end{cases}$$

and P a probability distribution on X, then:

$$HO(E,P) = H(P). \tag{14}$$

4 Observation degree as expected value of a random variable

In this section, a new interpretation of the observation degree is given. This degree was defined as:

$$\pi(x_j) = \sum_{x \in X} p(x) E(x, x_j).$$

When working on discrete domains, the T- indistinguishability operator can be represented by a symmetric matrix M, where the (i,j) component takes the value $E(x_i, x_j)$. A closer look to the columns (or rows) of matrix M, leads us to realize that column i contains the indistinguishability degrees of all the elements with respect to the element x_i. Therefore, we can define the fuzzy set "similarity with x_i" ($\approx_i$) on X, also called singletons or simply columns on the literature, as:

$$\forall x_j \in X \quad : \quad \approx_{x_i}(x_j) = E(x_i, x_j). \tag{15}$$

Now, we define the random variable $G_{\approx_{x_i}}$ over the interval [0,1] with the following probability distribution $P_{\approx_{x_i}}$:

$$\forall r \in [0,1] \quad : \quad P_{\approx_{x_i}}(r) = \sum_{x \in X | \approx_{x_i}(x) = r} p(x). \tag{16}$$

It is easy to see that $P_{\approx_{x_i}}$ is a probability distribution over the set of membership degrees ([0,1]), where each $r \in [0,1]$ takes as a probability, the sum of all the elements whose "similarity degree" with x_i is just r. Finally, the next equality holds:

$$\forall x \in X \quad : \quad \pi(x) = E(G_{\approx_x}). \tag{17}$$

, that is, the observation degree of an element x is the expected value of the random variable $G_{\approx_x}$. In other words, the observation degree is the expected value for the "similarity degrees with x".

5 Simultaneous observation degree

Arrived at this point, we will introduce the concept of simultaneous observation degree. The meaning of this new degree could be stated as follows: given that an indistinguishability relation has been defined, it is possible for two independent observers to disagree in the observation of an event. For instance, observer A may have observed event x_i, while observer B may have observed event x_j, if x_i and x_j are "close enough". Obviously, if the events were fully distinguishable, this "overlapping" or "simultaneous observation" could not have been possible (assuming the abssence of noise or error).

Definition 5.1: The simultaneous observation degree of the subset $\{x_1, \ldots x_k\}$ is defined by:

$$\begin{aligned} O(\{x_1, \ldots, x_k\}) &= E(G_{[\approx_{x_1} \wedge \ldots \wedge \approx_{x_k}]}) \\ &= E(G_{T(\approx_{x_1}, \ldots, \approx_{x_k})}). \end{aligned} \tag{18}$$

That is, the simultaneous observation degree of $\{x_1, \ldots, x_k\}$ is the expected value of the similarity degrees with respect to $x_1, \ldots, x_k$.

Lemma 5.2:

$$\begin{aligned} O(\{x_1, \ldots, x_k\}) &= 1 \leftrightarrow \\ \forall x_i, x_j \in X \quad &: \quad E(x_i, x_j) = 1 \end{aligned} \tag{19}$$

Lemma 5.3:

$$O(\{x_1, \ldots, x_k\}) = 0 \leftrightarrow$$
$$\forall x_i \in X \mid p(x_i) > 0 : \exists x_j \mid E(x_i, x_j) = 0 \quad (20)$$

6 Conditional observation degree

In the last section we dealt with the scene in which there was disagreement among observers "equipped" with the same indistinguishability relation. Now, we will examine the case of observers with different indistinguishability relations. Here, the main question is: How does affect the observations performed by an observer in order to predict the observations of some other observer?. For instance, let us suppose that we know that observer A using indistinguishability E_A has observed event x_i. This fact restricts the events that really might have been happened to the set of events "close enoug" to x_i with respect to E_A . This restriction in the set of possible events affects to the observability of observer B.

Definition 6.1: Let E, E' two T-indistinguishability operators on a set X, P a probability distribution on X and T a t-norm. The conditional observation degree of event x_j in (E, P), given that x_i has been observed in (E', P) is defined by:

$$\pi_{E|E'}(x_j \mid x_i) = \sum_{x \in X} p(x) T(E'(x_i, x), E(x, x_j)). \quad (21)$$

Lemma 6.2:

$$\pi_{E|E}(x_j \mid x_i) = O_E(\{x_i, x_j\}). \quad (22)$$

That is, when the T-indistinguishability operators are equal, the conditioned observation degree of x_j, given the observation x_i, is the simultaneous observation degree between x_i and x_j in E. It should be noted that the conditioned observation degree is not necessary symmetric, but when $E = E'$, symmetry holds.

7 Conclusions and future work

In this work we have studied the problem of adapting the notion of entropy when, besides the probabilistic uncertainty, we have the uncertainty coming from a indistinguishability relation defined over the set of events. New interpretations are given, which allow us to define the "observe" paradigm. This paradigm leads to the definition of some "observation" degrees, and to study their relationship with the concept of entropy. As future work we propose the next points:

- Joint observational entropy and conditional observational entropy.
- Extend the results to continuous domains .
- Application to the construction of fuzzy decision trees and clustering.

Acknowledgements

This work has been funded by the project PB94-1208 of the DGICYT.

References

[1] Fazlollah Reza. An introduction to information theory. *Dover Publications, inc. New York.*

[2] Pedro Gil (1981) Álvarez. Teoría matemática de la información. *Ediciones ICE.*

[3] J. Jacas and J. Recasens (1995). Fuzzy T-transitive relations: eigenvectors and generators. *Fuzzy sets and systems 72.* pp 147-154.

[4] Nikhil R. Pal and James C. Bezdek (1994). Measuring fuzzy uncertainty. *IEEE Transactions on Fuzzy Systems, Vol 2, No 2* pp 107-118.

[5] Ronald R. Yager (1992). Entropy measures under similarity relations. *Int. J. General Systems, Vol 20.* pp 341-358.

[6] L. Zadeh (1968). Probability measures of fuzzy events. *J. Math. Analysis and Applicat. Vol 23.* pp 421-427.