Entropy and monotonicity in artificial intelligence

Bernadette Bouchon-Meunier, Christophe Marsala

Entropy and Monotonicity in Artificial Intelligence

Bernadette Bouchon-Meunier

Sorbonne Université, CNRS, LIP6,
F-75005 Paris, France
Bernadette.Bouchon-Meunier@lip6.fr

Christophe Marsala

Sorbonne Université, CNRS, LIP6,
F-75005 Paris, France
Christophe.Marsala@lip6.fr

Abstract

Entropies and measures of information are extensively used in several domains and applications in Artificial Intelligence. Among the original quantities from Information theory and Probability theory, a lot of extensions have been introduced to take into account fuzzy sets, intuitionistic fuzzy sets and other representation models of uncertainty and imprecision. In this paper, we propose a study of the common property of monotonicity of such measures with regard to a refinement of information, showing that the main differences between these quantities come from the diversity of orders defining such a refinement. Our aim is to propose a clarification of the concept of refinement of information and the underlying monotonicity, and to illustrate this paradigm by the utilisation of such measures in Artificial Intelligence.

Keywords: Entropy, Monotonicity, Measure of fuzziness, Intuitionistic entropy measure, Divergence. Artificial Intelligence.

1. Introduction

The concept of information is complex and corresponds to factual elements regarding an event or a situation, such as observations or news, as well as the amount of knowledge brought by these factual elements. A piece of information

can be analysed either from the point of view of its content, semantics or meaning or from the point of view of its form, encoding or label. Claude Shannon [1] and Norbert Wiener [2] simultaneously introduced the first measure to evaluate information in 1948 in a general theory of communication. Both of them were only considering the information communicated by the observation of the occurrence of a message or an event among a set of messages or events, respectively in communications systems and in the framework of cybernetics. The numerical measure they introduced was based on probabilities and had nothing to do with semantics. In reaction to this proposal, Rudolf Carnap and Yehoshua Bar-Hillel [3] proposed to formalise the concept of semantic information on the basis of a logical approach of natural language, through the so-called amount of information. Later on, Edwin Thompson Jaynes [4] established the maximum entropy principle to make inferences on the basis of partial information, by looking for the maximum entropy, given the available knowledge. Solomon Kullback then introduced with Richard Leibler and developed the concept of discrimination measure or divergence enabling to compare two probability distributions [5, 6] .

Artificial intelligence was born in 1956 during the well-known "Dartmouth Summer Research Project on Artificial Intelligence" [7]. Even though it was not a direct emanation of cybernetics, it was strongly interrelated with it and it would have looked natural to consider Wiener's measure of information in Artificial Intelligence. It was not the case because Artificial Intelligence originally limited itself to symbolic methods, based on classical logic, rejecting any kind of numerical treatment of information, including probabilities. It might have looked relevant to use Carnap-Bar Hillel's measure of information, but this theory has never been applied to Artificial Intelligence, because of its numerical nature.

It is only in the 80s that some branches of Artificial Intelligence established a bridge with information theory. The Kullback discrimination measure [8] and the maximum entropy principle [9] were for instance investigated in the framework of Bayesian inference and updating. One of the first consistent utilisations

of measures of information in machine learning was Quinlan's decision trees [10], based on a concept of information gain used to choose an attribute providing its maximum value [4]. At the same time, several other seminal works changed the view of numerical methods in Artificial Intelligence. Nilsson's probabilistic logic [11] promoted Shannon's measure of information to develop a method to implement probabilistic entailment. Judea Pearl's early works and his book [12] largely promoted the use of probabilities in Artificial Intelligence, his belief networks being considered as semantics-based systems. The optimisation of an entropy or a gain of information they recommended became popular in logic (for instance [13]) before being more largely applied in Artificial Intelligence.

Since this period, measuring the information provided by the observation of events has been a challenge. A number of quantities has been pointed out and studied in the literature to achieve this goal, called entropies or measures of information in the original probabilistic framework. Their extension to other frameworks such as fuzzy knowledge representation or its generalisations have given rise to the study of other aspects of information, for instance fuzziness or specificity. They have often been constructed by analogy with the probabilistic case, which may look artificial, but their properties go far beyond a simple analogy. They refer to the information contained in belief functions, for instance in [14, 15], or fuzzy systems, for instance in [16]. Such entropies are used in Bayesian networks [17] or decision making systems [18, 19]. We propose to continue the analysis initiated in [20] and formalised in [21], in focusing on the property of monotonicity, which appears essential for the notion of entropy and its utilisation in Artificial Intelligence. Most of the measures existing in the literature point out properties which have in common to be related to different aspects of monotonicity based on an order taking into account the context, the point of view or the knowledge representation. The entropy increases when the means to perform observations is refined, the concept of refinement taking various forms, from the utilisation of a better tool providing less vagueness to the utilisation of additional features to obtain more detailed information or, to express it in a more colourful language in the case where a photo brings

information on an object, from zooming on the object to taking additional photos from another side of the object. The importance of the monotonicity of entropy, considered under the angle of the coarseness of information, has been pointed out in [22]. Other views of the monotonicity of entropy have been proposed in random variable analysis, see for instance [23].

This paper does not pretend to review all entropies or quantities of information introduced in the literature and related to artificial intelligence, in any way, as this would be far beyond the size of this manuscript. We only focus on a study of the common property of monotonicity of such measures with regard to a refinement of information, showing that the main differences between these quantities come from the diversity of orders defining such a refinement. Our aim is to propose a clarification of the concept of refinement of information and the underlying monotonicity, and to illustrate this paradigm by the utilisation of such measures in Artificial Intelligence.

The paper is organised as follows. In Section 2, we introduce entropy measures and definitions of monotonicity according to three different forms associated with different visions of the refinement of information. In Section 3, we illustrate these visions on various classic entropy measures: probabilistic entropies, measures of fuzziness, similarity relation-based entropy measures, entropies in the settings of the Atanassov intuitionistic fuzzy sets, and we develop also a study on popular divergence measures. In Section 4, we review some common uses of entropies in Artificial Intelligence and their highlighted properties of monotonicity. Finally, some conclusions are drawn and a set of perspectives is presented in Section 5.

2. Monotonicity of entropy measures

Shannon [1] clearly based his definition of entropy, considered as "*measures of information, choice and uncertainty*" on a concept of monotonicity, as he states "*The uncertainty of y is never increased by knowledge of x. It will be decreased unless x and y are independent events, in which case it is not changed*",

which can be regarded as an interpretation of recursivity he furthermore requires from entropy. Additivity and recursivity are among the most important algebraic properties of entropy [24], and they imply an increase of the entropy resulting from the refinement of information acquired on an event through observations. Later on, Renyi [25] introduced the first of a long list of generalisations of Shannon entropy, still satisfying a property of additivity. It is worth mentioning Mugur-Schächter's work [26] on the general relativity of descriptions. She considers that any process of knowledge extraction is associated with epistemic operators called a delimiter and a view, representing the influence of the context and the observation tool on the considered event.

A refinement of information results from a change in the observation tool. In his generalised information theory, Kampé de Fériet [27, 28] takes into account observers and also requests a monotonicity of information with respect to an order on events.

Information theory do not pretend to evaluate all aspects, and it provides an evaluation of the decrease of uncertainty after an observation of events by means of entropies.

Given the amount of data available in the numerical world, which is covering all aspects of modern life, evaluating information is a major issue. All tools enabling the user to compare two pieces of information, to evaluate the information available in a given environment, to make diagnosis or prediction on the basis of information provided by observations or data, to aggregate chunks of information, are useful. Unfortunately, there are many such tools and it is difficult to see their common features. This is why we propose to analyse measures of information and to revisit classic approaches of information evaluation in order to focus on monotonicity which we consider the most natural and relevant property requested from such a tool.

Let us consider a set of objects or events that represent the real world. In this paper, for the sake of simplicity, we only consider finite sets, but this work could be generalised to non-countable sets. We use the notation proposed in the seminal paper by Aczél and Daróczy on the so-called inset entropy [29] to

formalise the available information on the set of objects or events and taking into account the context, the point of view and the chosen knowledge representation.

We consider an algebra $\mathcal{B}$ defined on a finite universe $\mathcal{U}$. For any integer $n > 0$, we note:

- $X_n = \{(x_1, \ldots, x_n) \mid x_i \in \mathcal{B},\ x_i \cap x_j = 0 \text{ if } i \neq j,\ \forall i,\ j = 1,\ \ldots, n\}$;
- $P_n = \{(p_1, \ldots, p_n) \mid p_i \in [0, 1]\}$, p_i being associated with x_i through a function $p : \mathcal{B} \to [0, 1]$, a particular case being a probability measure defined on $(\mathcal{U}, \mathcal{B})$;
- $W_n = \{(w_{x_1}, \ldots, w_{x_n}) \mid w_{x_i} \in \mathcal{R}^+, \forall i = 1, \ldots, n\}$, a family of n-tuples of weights[1] associated with n-tuples of $\mathcal{B}$ through a function $f : \mathcal{B} \to \mathcal{R}^{+n}$, such that $f(x_1, \ldots, x_n) = (w_{x_1}, \ldots, w_{x_n})$.

Similarly to the definition of inset entropy [29], we introduce an **entropy measure** as a sequence of mappings $E_n : X_n \times P_n \times W_n \to R^+$ satisfying several properties among a long list, for instance available in [24] or in [30].

In the sequel, for the sake of simplicity, we use the notation

$$\begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ w_{x_1}, & \ldots, & w_{x_n} \end{pmatrix}$$

rather than $((x_1, \ldots, x_n), (p_1, \ldots, p_n), (w_{x_1}, \ldots, w_{x_n}))$ to represent an element of $X_n \times P_n \times W_n$, according to the notation used in [29].

We claim that the most significant properties to characterise an entropy measure are relative to *monotonicity* with respect to a refinement of information which can take various forms, depending on a chosen order. Such a monotonicity corresponds to the natural idea that the more details, precision, certainty we obtain from the observation of objects or events, or equivalently the more refined

[1] In the following, for the sake of simplicity, w_{x_i} is denoted w_i when the meaning of i is clear.

information we have on them, the bigger the amount of information we have on these objects or events.

To use a metaphor, we can consider that we are facing a picture of an object providing some amount of information on it. We can first improve the light on the object before taking another picture in order to decrease the fuzziness of the details, or take another picture with a higher resolution, which gives more information on the object according to an increase of the precision, both cases corresponding to a monotonicity described in 2.1. We can also select a part of the object and make several pictures of this part, in a form of weak recursivity described in 2.2. We can finally partition the object into different parts and, for each of them, make more pictures providing a bigger amount of information on the object, in a form of weak additivity presented in 2.3.

We present these three forms of monotonicity which can be adapted to the knowledge representation we choose, as highlighted in the next sections.

2.1. O-monotonicity

The first form of monotonicity, noted **O-monotonicity**, is defined according to a given (partial) order on the elements: the monotonicity highlights a link between the order of the elements and the order induced by the measure.

2.1.1. Definition

Let $\prec$ be a **partial order** on a reliable observation of the objects or events, O-monotonicity can be written as follows: if

$$\begin{pmatrix} x_1, & \dots, & x_n \\ p_1, & \dots, & p_n \\ w_1, & \dots, & w_n \end{pmatrix} \prec \begin{pmatrix} x'_1, & \dots, & x'_n \\ p'_1, & \dots, & p'_n \\ w'_1, & \dots, & w'_n \end{pmatrix}$$

then

$$E_n \begin{pmatrix} x_1, & \dots, & x_n \\ p_1, & \dots, & p_n \\ w_1, & \dots, & w_n \end{pmatrix} \leq E_n \begin{pmatrix} x'_1, & \dots, & x'_n \\ p'_1, & \dots, & p'_n \\ w'_1, & \dots, & w'_n \end{pmatrix}$$

2.1.2. Particular cases

Examples of monotonicity can be based on the following partial orders:

$$\begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ w_1, & \ldots, & w_n \end{pmatrix} \prec \begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ w'_1, & \ldots, & w'_n \end{pmatrix}$$

if and only if one of the following conditions is satisfied: $(O1)$, called **sharpness** in [31], or $(O2)$

(O1) $\forall i = 1, \ldots, n$, if $w'_i \geq \frac{1}{2}$ then $w'_i \leq w_i$;

(O2) $\forall i = 1, \ldots, n,\ w'_i \geq w_i$.

These examples are based on an order related to W. Other orders, for instance related to P, could be also used. Such examples will be studied later.

2.2. R-monotonicity

The second form of monotonicity, noted **R-monotonicity**, is based on a decrease of the coarseness of a partition of the universe of discourse.

2.2.1. Definition

R-monotonicity can correspond to a property of **weak recursivity** defined as follows:

$$E_n \begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ w_{x_1}, & w_{x_2}, & \ldots, & w_{x_n} \end{pmatrix} \geq E_{n-1} \begin{pmatrix} x_1 \cup x_2, & x_3, & \ldots, & x_n \\ p_1 + p_2, & p_3, & \ldots, & p_n \\ w_{x_1 \cup x_2} & w_{x_3}, & \ldots, & w_{x_n} \end{pmatrix}$$

A particular case of weak recursivity is what we call the ψ-recursivity, defined

for a function $\psi : X_2 \times P_2 \times W_2 \to R^+$ as:

$$E_n \begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ w_{x_1}, & w_{x_2} & \ldots, & w_{x_n} \end{pmatrix} =$$

$$E_{n-1} \begin{pmatrix} x_1 \cup x_2, & x_3 & \ldots, & x_n \\ p_1 + p_2, & p_3 & \ldots, & p_n \\ w_{x_1 \cup x_2}, & w_{x_3} & \ldots, & w_{x_n} \end{pmatrix} + \psi \begin{pmatrix} x_1, & x_2 \\ p_1, & p_2 \\ w_{x_1}, & w_{x_2} \end{pmatrix} E_2 \begin{pmatrix} x_1, & x_2 \\ \frac{p_1}{p_1+p_2}, & \frac{p_2}{p_1+p_2} \\ w_{x_1}, & w_{x_2} \end{pmatrix}.$$

2.2.2. Particular case

The classic property of **recursivity** corresponds to:

$$\psi_0 \begin{pmatrix} x_1, & x_2 \\ p_1, & p_2 \\ w_{x_1}, & w_{x_2} \end{pmatrix} = p_1 + p_2,$$

where the weights are not taken into account.

2.3. A-monotonicity

The third form of monotonicity, noted **A-monotonicity**, is based on the consideration of a secondary finite universe $\mathcal{U}'$ and an algebra $\mathcal{B}'$ on $\mathcal{U}'$ providing more details on the observed phenomenon or object, through additional observations.

2.3.1. Definition

Similarly to the situation on $\mathcal{U}$, we consider for any integer m

- $X'_m = \{(x'_1, \ldots, x'_m) \mid x'_i \in \mathcal{B}',\ \forall i\}$;
- $P'_m = \{(p'_1, \ldots, p'_m) \mid p'_i \in [0,1]\}$, p'_i being associated with x'_i through a function $p' : \mathcal{B}' \to [0,1]$;
- $W'_m = \{(w'_1, \ldots, w'_m) \mid w'_i \in \mathcal{R}^+; \forall i\}$, a family of m-tuples of weights associated with m-tuples of elements of $\mathcal{B}'$ through a function $f' : \mathcal{B}' \to \mathcal{R}^+$, such that $f'(x'_i) = w'_i$.

We further suppose that there exist two combination operators $\star$ and $\circ$ enabling us to equip the Cartesian product of $\mathcal{U} \times \mathcal{U}'$ with similar distributions:

- $P_n \star P'_m = \{(p_1 \star p'_1, \ldots, p_i \star p'_j, \ldots) \mid p_i \star p'_j \in [0,1]\}$, $p_i \star p'_j$ being associated with (x_i, x'_j) for any i and j through a function $p \star p'$,
- $W_n \circ W'_m = \{(w_{1,1}, \ldots, w_{i,j}, \ldots) \mid w_{i,j} \in \mathcal{R}^+, \ \forall i, j\}$, is defined through a function $f \circ f' : \mathcal{B} \times \mathcal{B}' \to \mathcal{R}^+$, such that: $f \circ f'(x_1, x'_j) = w_{i,j}$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

Such a refinement leads to a property of **weak additivity** stating the following:

$$E_{n\times m}\begin{pmatrix} (x_1,x'_1), & (x_1,x'_2), & \ldots, & (x_i,x'_j), & \ldots, & (x_n,x'_m) \\ p_1 \star p'_1, & p_1 \star p'_2, & \ldots, & p_i \star p'_j, & \ldots, & p_n \star p'_m \\ w_{x_1,x'_1}, & w_{x_1,x'_2}, & \ldots, & w_{x_i,x'_j}, & \ldots, & w_{x_n,x'_m} \end{pmatrix} \geq \max\left[E_n\begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ w_1, & \ldots, & w_n \end{pmatrix}, E_m\begin{pmatrix} x'_1, & \ldots, & x'_m \\ p'_1, & \ldots, & p'_m \\ w'_1, & \ldots, & w'_m \end{pmatrix} \right]$$

2.3.2. Particular case

The classic **additivity** property stands in the case where $\mathcal{U}$ and $\mathcal{U}'$ are independent universes, p and p' being probability distributions on $(\mathcal{U}, \mathcal{B})$ and $(\mathcal{U}', \mathcal{B}')$, weights generally not being taken into account. It yields:

$$E_{n\times m}\begin{pmatrix} (x_1,x'_1), & (x_1,x'_2), & \ldots, & (x_i,x'_j), & \ldots, & (x_n,x'_m) \\ p_1 \star p'_1, & p_1 \star p'_2, & \ldots, & p_i \star p'_j, & \ldots, & p_n \star p'_m \\ w_{x_1,x'_1}, & w_{x_1,x'_2}, & \ldots, & w_{x_i,x'_j}, & \ldots, & w_{x_n,x'_m} \end{pmatrix} = E_n\begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ w_{x_1}, & \ldots, & w_{x_n} \end{pmatrix} + E_m\begin{pmatrix} x'_1, & \ldots, & x'_m \\ p'_1, & \ldots, & p'_m \\ w_{x'_1}, & \ldots, & w_{x'_m} \end{pmatrix}$$

3. Diverse entropy measures

In this section, the previous definitions of monotonicity are studied in the case of classical and well-known entropy measures.

3.1. Shannon and weighted entropies

It is well-known that the classic **Shannon entropy** defined as:

$$E_n^S(p) = -\sum_{i=1}^{n} p_i \log p_i,$$

only taking into account X_n and P_n, is additive and recursive and then R-monotonous and A-monotonous.

Its generalisation to the case where weights are associated with events to represent a cost or an importance is a **weighted entropy** defined on $X_n \times P_n \times W_n$ as follows [32]:

$$E_n^w \begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ w_1, & w_2, & \ldots, & w_n \end{pmatrix} = -\sum_{i=1}^{n} w_i \, p_i \log p_i.$$

The weighted entropy is O-monotonous with regard to the partial order (O2). It is also recursive, and then R-monotonous when considering

$$w_{x_1 \cup x_2} = \frac{(p_1 w_1 + p_2 w_2)}{(p_1 + p_2)}.$$

In addition, the weighted entropy is A-monotonous as soon as we consider an aggregation function at least equal to the maximum:

$$E_{n\times m} \begin{pmatrix} (x_1, x'_1), & \ldots & \ldots, & (x_n, x'_m) \\ p_1 \star p'_1, & \ldots & \ldots, & p_n \star p'_m \\ w_1 \circ w'_1, & w_1 \circ w'_2, & \ldots, & w_n \circ w'_m \end{pmatrix} \geq$$

$$\max \left[E_n \begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ w_1, & \ldots, & w_n \end{pmatrix}, E_m \begin{pmatrix} x_1, & \ldots, & x_m \\ p_1, & \ldots, & p_m \\ w'_1, & \ldots, & w'_m \end{pmatrix} \right]$$

whenever $w_i \circ w'_j \geq max(w_i, w'_j)$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$

An entropy of the same form as the weighted entropy has been introduced by Zadeh [33] in the case where weights are replaced by membership degrees, and the so-called *entropy of a fuzzy set* is defined as:

$$E_n^Z \begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ \mu_1, & \mu_2, & \ldots, & \mu_n \end{pmatrix} = -\sum_{i=1}^{n} \mu_i \, p_i \log p_i.$$

It is obviously still O-monotonous (with partial order (O2)), and A-monotonous according to [33].

3.2. Parameterised entropies

A number of generalisations of Shannon's entropies flourished in the 60s and in the 70s, independent of W_n and preserving some of the basic properties of Shannon's entropy. The first one was the Renyi's entropy of order α, defined as:

$$E_n^{R\alpha} \begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ w_1, & w_2, & \ldots, & w_n \end{pmatrix} = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} {p_i}^{\alpha},$$

for a parameter α strictly positive and different from 1, the Shannon entropy corresponding to the limit case when α tends to 1. This quantity is additive, as proved in [25], and then A-monotonous, but not recursive. It is easy to see that Renyi's entropy of order α is nevertheless R-monotonous.

Another parameterised form of entropy is Daróczy's entropy of type β, for a parameter β strictly positive and different from 1, defined as [34]:

$$E_n^{D\beta}(p) = \frac{1}{2^{1-\beta} - 1} (\sum_{i=1}^{n} {p_i}^{\beta} - 1).$$

The Shannon entropy corresponds to the limit case when β tends to 1. When $\beta = 2$, we obtain a quantity proportional by a factor 2 to the Gini diversity index [35] used in the construction of decision trees by the Cart method.

It was known not to be either additive or recursive and was proved to satisfy recursivity of type β, equivalent to the weak recursivity we consider, with:

$$\psi \begin{pmatrix} x_1, & x_2 \\ p_1, & p_2 \\ w_{x_1}, & w_{x_2} \end{pmatrix} = (p_1 + p_2)^\beta.$$

It also satisfies an additivity of type β, implying the weak additivity. The Daróczy's entropy of type β is then R-monotonous and A-monotonous.

3.3. Measure of fuzziness

Shortly after the weighted entropy, another entropy measure was introduced by De Luca and Termini by analogy with the Shannon entropy, but in a non-probabilistic framework [31], and then independently of P_n. It is a measure of fuzziness, in the case where f is the membership function of a fuzzy set on U:

$$E_n^{DLT} \begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ w_1, & w_2, & \ldots, & w_n \end{pmatrix} = -\sum_{i=1}^{n} w_i \log w_i - \sum_{i=1}^{n} (1 - w_i) \log(1 - w_i).$$

A major property of this quantity is its O-monotonicity with respect to the above mentioned partial order *(O1)* defining the sharpness.

It can further be observed that, in the case where the weights are possibility degrees, *ie.* $\max(w_1, \ldots, w_n) = 1$, this measure of fuzziness is also weakly recursive and then R-monotonous:

$$E_n^{DLT} \begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ w_1, & w_2, & \ldots, & w_n \end{pmatrix} \geq E_{n-1}^{DLT} \begin{pmatrix} x_1 \cup x_2, & x_3, & \ldots, & x_n \\ p_1 + p_2, & p_3, & \ldots, & p_n \\ \max(w_1, w_2), & w_3, & \ldots, & w_n \end{pmatrix}$$

3.4. Entropy measures under similarity relations

We consider a similarity relation S on $U = \{x_1, \ldots, x_n\}$, reflexive, symmetric and min-transitive. R.R. Yager [36] defines the following entropy measure on $X_n \times P_n \times W_n$:

$$E_n^{Sim} \begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ \overline{S}_1, & \overline{S}_2, & \ldots, & \overline{S}_n \end{pmatrix} = -\sum_{x_i \in U} p_i \log \overline{S}_i$$

with $\overline{S}_i = \sum_{x_j \in U} p_j \, S(x_i, x_j)$ for all $i = 1, \ldots n$.

The similarity reflects a point of view on the n events, making explicit to which extent they are similar with regard to a given criterion. If we consider two different points of view, symbolised by two similarity relations S and S', we can show that this entropy measure is O-monotonous with respect to the order $(O3)$:

$$\begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ \overline{S}_1, & \ldots, & \overline{S}_n \end{pmatrix} \leq \begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ \overline{S}'_1, & \ldots, & \overline{S}'_n \end{pmatrix}$$

if and only if similarities S and S' satisfy:

(O3) $S \preceq S' \Leftrightarrow S(x_i, x_j) \leq S'(x_i, x_j) \; \forall i, j.$

This entropy measure is also A-monotonous, if we define a joint similarity relation $S \times S'$ on the Cartesian product $U \times U'$ as follows, for two similarity relations S defined on U and S' defined on U':

$$S \times S'((x_i, y_j), (x_k, y_l)) = \min\left(S(x_i, x_k), S'(y_j, y_l)\right)$$

for any x_i and x_k in U, any y_j and y_l in U'.

3.5. Ambiguity or nonspecificity

In the framework of possibility distributions, corresponding to a fuzzy set-based knowledge representation in which a membership degree is interpreted as the possibility of the observed variable to take a given value, with a maximum equal to 1, a measure of ambiguity or non-specificity has been introduced by [37] and called U-uncertainty. It is defined under the hypothesis that the x_i are ranked according to a possibility distribution: $\pi_1 \geq \pi_2 \geq ... \geq \pi_n$ and $\pi_1 > 0$.

This quantity, independent of P, got more popularity when pointed out by [38] in the induction of fuzzy decision trees.

$$E_n^{HK}\begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ \pi_1, & \pi_2, & \ldots, & \pi_n \end{pmatrix} = \sum_{i=1}^{n-1}(\pi_i - \pi_{i+1}) \log i.$$

It was proved [39] that this quantity is O2-monotonous. It is also additive, then A-monotonous, when the two possibility distributions are non-interactive, as follows:

$$E_{n\times m}^{HK}\begin{pmatrix} (x_1, x'_1), & \ldots, & (x_n, x'_m) \\ p_1 \star p'_1, & \ldots, & p_n \star p'_m \\ min(\pi_1, \pi'_1), & \ldots, & min(\pi_n, \pi'_m) \end{pmatrix} \geq$$
$$\max\left[E_n^{HK}\begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ \pi_1, & \ldots, & \pi_n \end{pmatrix}, E_m^{HK}\begin{pmatrix} x_1, & \ldots, & x_m \\ p_1, & \ldots, & p_m \\ \pi'_1, & \ldots, & \pi'_m \end{pmatrix}\right]$$

It is also recursive, and then R-monotonous.

3.6. Divergence

Kulback and Leibler's divergence [5, 6] is another form of entropy, when W_n is identical with P_n, corresponding to the following:

$$J_n^{KL}\begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ q_1, & q_2, & \ldots, & q_n \end{pmatrix} = \sum_{i=1}^{n} p_i \log(\frac{p_i}{q_i}),$$

where $p_1, p_2, \ldots, p_n$ and $q_1, q_2, \ldots, q_n$ are two probability distributions on the same set of events, for instance a prior and a posterior distribution. They prove that J_n^{KL} is additive, and therefore A-monotonous. This divergence is also

recursive, such that:

$$J_n^{KL}\begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ q_1, & q_2 & \ldots, & q_n \end{pmatrix} =$$

$$J_{n-1}^{KL}\begin{pmatrix} x_1 \cup x_2, & x_3 & \ldots, & x_n \\ p_1 + p_2, & p_3 & \ldots, & p_n \\ q_1 + q_2, & q_3 & \ldots, & q_n \\ & & & \end{pmatrix} + (p_1 + p_2) J_2^{KL}\begin{pmatrix} x_1, & x_2 \\ \frac{p_1}{p_1+p_2}, & \frac{p_2}{p_1+p_2} \\ \frac{q_1}{q_1+q_2}, & \frac{q_2}{q_1+q_2} \end{pmatrix}$$

It is therefore R-monotonous.

Parameterised forms of divergence were introduced, in the same spirit as the parameterised entropies. The first one is divergence of order α introduced by Renyi [25] as the gain of information resulting from the replacement of $q_1, q_2, \ldots, q_n$ by $p_1, p_2, \ldots, p_n$, defined as follows:

$$J_n^{R\alpha}\begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ q_1, & q_2, & \ldots, & q_n \end{pmatrix} = \frac{1}{\alpha - 1} \log \sum_{i=1}^{n} \frac{p_i^\alpha}{q_i^{\alpha-1}},$$

for $\alpha \geq 0$ and $\alpha \neq 1$. It is additive, and therefore A-monotonous.

The second parameterised divergence was introduced by Rathie and Kannappan [40] as the directed divergence of type β:

$$J_n^{RK\beta}\begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2, & \ldots, & p_n \\ q_1, & q_2, & \ldots, & q_n \end{pmatrix} = \frac{1}{2^{\beta-1} - 1} (\sum_{i=1}^{n} \frac{p_i^\beta}{q_i^{\beta-1}} - 1).$$

The authors proved that it is R-monotonous. Furthermore, they prove that it has a form of strong non-additivity entailing the following:

$$J_{n\times m}^{RK\beta}\begin{pmatrix} (x_1,x'_1), & (x_1,x'_2), & \ldots, & (x_i,x'_j), & \ldots, & (x_n,x'_m) \\ p_1\star p'_1, & p_1\star p'_2, & \ldots, & p_i\star p'_j, & \ldots, & p_n\star p'_m \\ q_1\star q'_1, & q_1\star q'_2, & \ldots, & q_i\star q'_j, & \ldots, & q_n\star q'_m \end{pmatrix} =$$

$$J_n^{RK\beta}\begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ q_1, & \ldots, & q_n \end{pmatrix} + \sum_{i=1}^{n}\frac{p_i^{\beta}}{q_i^{1-\beta}}J_m^{RK\beta}\begin{pmatrix} x'_1, & \ldots, & x'_m \\ p'_1, & \ldots, & p'_m \\ q'_1, & \ldots, & q'_m \end{pmatrix}$$

which implies a property of A-monotonicity

3.7. Intuitionistic entropy measures

In this section, we consider the setting of the Atanassov intuitionistic fuzzy sets (AIFS) where several entropy measures have been introduced [41, 42]. First of all, some basics of AIFS are recalled.

Let X be a universe, an *Atanassov intuitionistic fuzzy set* (AIFS) A of X is defined [43] by:

$$A = \{(x, \mu_A(x), \nu_A(x)) | x \in X\}$$

with $\mu : X \to [0,1]$, $\nu : X \to [0,1]$ and $0 \leq \mu_A(x) + \nu_A(x) \leq 1$, $\forall x \in X$. Here, $\mu_A(x)$ and $\nu_A(x)$ represent respectively the membership degree and the non-membership degree of x in A. Given an intuitionistic fuzzy set A of X, the hesitancy lying on the membership of x to A is the *intuitionistic index of x to A* defined for all $x \in X$ as $\pi_A(x) = 1 - (\mu_A(x) + \nu_A(x))$. It is easy to see that we always have $\pi_A(x) \in [0,1]$.

The inclusion of AIFS is defined as: $A \subseteq B$ if and only if $\mu_A(x) \leq \mu_B(x)$ and $\nu_A(x) \geq \nu_B(x)$, $\forall x \in X$.

The union of two AIFS A and B is defined as the AIFS $A \cup B$ such that $\mu_{A\cup B}(x) = \max(\mu_A(x), \mu_B(x))$ and $\nu_{A\cup B}(x) = \min(\nu_A(x), \nu_B(x))$, $\forall x \in X$. The intersection of two AIFS A and B is defined as the AIFS $A \cap B$ such that $\mu_{A\cap B}(x) = \min(\mu_A(x), \mu_B(x))$ and $\nu_{A\cap B}(x) = \max(\nu_A(x), \nu_B(x))$, $\forall x \in X$. It can be easily seen that $\mu_{A\cup B}(x) \in [0,1]$ and $\mu_{A\cap B}(x) \in [0,1]$.

3.7.1. Definitions of entropy measures for AIFS.

Several works in AIFS theory have proposed the definition for an *entropy of an intuitionistic fuzzy set A*. With our notations, we represent these quantities as:

$$E_n^{IFS}\begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ (\mu_1,\nu_1), & \ldots, & (\mu_n,\nu_n) \end{pmatrix},$$

with $\mu_A(x_i) = \mu_i$ and $\nu_A(x_i) = \nu_i$. Here, the weights from W_n are thus defined as tuples from $[0,1] \times [0,1]$.

There exist various definitions of entropy measures in the AIFS setting [41], independent of P_n. For instance, the entropy measure given in [42] is defined as:

$$E_n^{IFS,S}\begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ (\mu_1,\nu_1), & \ldots, & (\mu_n,\nu_n) \end{pmatrix} = 1 - \frac{1}{2n}\sum_{i=1}^{n} |\mu_i - \nu_i|$$

In [41], the following entropy measure is also introduced:

$$E_n^{IFS,G}\begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ (\mu_1,\nu_1), & \ldots, & (\mu_n,\nu_n) \end{pmatrix} = \frac{1}{2n}\sum_{i=1}^{n} \big(1 - |\mu_i - \nu_i|\big)(1+\pi_i),$$

with $\pi_i = 1 - (\mu_i + \nu_i)$.

Another way to define an entropy measure is presented in [44] where the definition is based on extensions of the Hamming distance and the Euclidean distance to AIFS. For instance, the following entropy measure is proposed:

$$E_n^{IFS,B}\begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ (\mu_1,\nu_1), & \ldots, & (\mu_n,\nu_n) \end{pmatrix} = \sum_{i=1}^{n} \pi_i$$

3.7.2. Entropy measures for AIFS and monotonicity.

In [41], it is recalled that, in the AIFS setting, a monotonicity property for an entropy measure could be ensured by definition. The authors present several

definitions that lie on the definition of a partial order on W_n and the concept of *less fuzzy than*. The following definitions of partial order could be used.

(O4) A is *less fuzzy* than B if for all $x \in X$
$\mu_A(x) \leq \mu_B(x)$ and $\nu_A(x) \geq \nu_B(x)$ if $\mu_B(x) \leq \nu_B(x)$,
or $\mu_A(x) \geq \mu_B(x)$ and $\nu_A(x) \leq \nu_B(x)$ if $\mu_B(x) \geq \nu_B(x)$,

and

(O5) A is *less fuzzy* than B if
$\mu_A(x) \leq \mu_B(x)$ and $\nu_A(x) \leq \nu_B(x), \forall x \in X$,

O-monotonicity. It is easy to see that *(O4)* and *(O5)* yield two versions of O-monotonicity. These definitions of monotonicity produce particular forms of E^{IFS}:

- $E_n^{IFS,S}$ satisfies the monotonicity based on *(O4)* as it is stated in [42];
- $E_n^{IFS,G}$ satisfies the monotonicity based on *(O4)*, as it is stated in [41];
- $E_n^{IFS,B}$ satisfies the monotonicity based on *(O5)* as it could be found in [44] where this entropy is given as example.

R-monotonicity. The measure $E_n^{IFS,B}$ satisfies the R-monotonicity if we have

$$E_n^{IFS,B}\begin{pmatrix} x_1, & \ldots, & x_n \\ p_1, & \ldots, & p_n \\ (\mu_1,\nu_1), & \ldots, & (\mu_n,\nu_n) \end{pmatrix} \geq$$
$$E_{n-1}^{IFS,B}\begin{pmatrix} x_1 \cup x_2, & x_3, & \ldots, & x_n \\ p_1 + p_2, & p_3, & \ldots, & p_n \\ (\max(\mu_1,\mu_2), \min(\nu_1,\nu_2)), & (\mu_3,\nu_3), & \ldots, & (\mu_n,\nu_n) \end{pmatrix}.$$

considering the union of AIFS as defined in the introduction of this section.

Hereafter, for the sake of simplicity, we note these two measures $E_n^{IFS,B}$ and $E_{n-1}^{IFS,B}$ respectively.

We have

$$\begin{aligned} E_n^{IFS,B} - E_{n-1}^{IFS,B} &= 1 - \mu_1 - \nu_1 + 1 - \mu_2 - \nu_2 - 1 + \max(\mu_1, \mu_2) + \min(\nu_1, \nu_2) \\ &= 1 + (\max(\mu_1, \mu_2) - \mu_1 - \mu_2) + (\min(\nu_1, \nu_2) - \nu_1 - \nu_2) \end{aligned}$$

and thus

$$E_n^{IFS,B} - E_{n-1}^{IFSB} = 1 - \min(\mu_1, \mu_2) - \max(\nu_1, \nu_2)$$

This corresponds to the intuitionistic index of the intersection of AIFS, and thus, as a consequence, we have $E_n^{IFS,B} - E_{n-1}^{IFS,B} \geq 0$ and $E_n^{IFS,B}$ satisfies the R-monotonicity.

4. Entropies in Artificial Intelligence

Entropies are very commonly used in Artificial Intelligence. Indeed, their monotonicity properties could be one of the main reasons for this success. In the following, some applications of entropies in Artificial Intelligence are presented and the kind of monotonicity involved is highlighted (that could explain why such an entropy is chosen in such applications). Our aim is not to propose a complete review of such applications but to show that an entropy can be considered as a universal tool.

Beyond the search of the maximum entropy common when probabilistic measures are used, monotonicity or maximisation of entropy has also been the core of methods based on non-probabilistic entropies.

4.1. Maximum Entropy Principle

A derived utilisation of the concept of monotonicity is the very commonly used *maximum entropy principle* (Maxent). It was first proposed by Jaynes [4], in the simple case where weights are not involved, as a way to choose the most appropriate probability distribution to cope with the uncertainty, as the one being "*maximally noncommittal with regard to missing information*". Jaynes introduced this principle in the case of the Shannon entropy, because of the easiness to solve the optimisation problem of maximising the entropy under

the condition of probabilities having a sum equal to 1. His aim was to benefit of a form of monotonicity of the entropy with respect to missing information tolerance. In his work, Jaynes argued [4] that "*Mathematically, the maximum-entropy distribution has the important property that no possibility is ignored*" and he based his proof on the fact that "*if all the p_i are equal, the quantity $A(n) = H(\frac{1}{n}, \ldots, \frac{1}{n})$ is a monotonic increasing function of n*".

In our setting, this Maxent principle can be related to the R-monotonicity of the Shannon entropy: the higher the number of non-null probabilities, the higher the entropy. It is easy to see this fact when considering a set of events on which a distribution of probability should be identified. Let P_1 and P_2 be two distributions of probability on this set of events. Between P_1 and P_2, Jaynes' principle argues that the distribution that the most covers the set of events should be preferred, which means the one that maximises the number of non-null probabilities associated with the events. In the case where P_1 and P_2 have the same number of non-null probabilities, the one that provides a more homogeneous distribution is preferred (the one that maximises the Shannon entropy). This property is, in fact, a side effect of the Jaynes' maximum entropy principle.

Since the R-monotonicity could be considered as the heart of this principle, it highlights the fact that choosing between two distributions by using Maxent, does not provide any specific information on the relative position of two distributions of probability but only a general information about their relative spread.

In the same spirit, and again in the case where the only available information is provided by probability distributions, Kullback [6] introduced the concept of *minimum discrimination information*, corresponding to the minimum value of the Kullback and Leibler's divergence.

These two principles have been extended to other entropies or divergences and widely used in Artificial Intelligence.

4.2. Entropies in Machine Learning

As previously said in the introduction, one of the well-known uses of entropies is in machine learning where the Shannon entropy is very popular.

In supervised learning, a common use of such a measure is dedicated to the learning of a decision tree from a training data set [10], [35]. Shannon entropy is not the unique measure to be used in such a process [45] but it is one of the most efficient. We can for instance also cite the Gini index of diversity which is also popular [35].

Here, for this kind of tree-like splitting processes, the R-monotonicity is the main property that is sought for. Indeed, measures following this kind of monotonicity property enable the better choice of description attributes when splitting the training set to reduce the uncertainty related to the prediction of the class.

In the building of fuzzy decision trees, other measures have been introduced. For instance, Renyi's entropy was used in presence of unbalanced datasets in [46]. In a fuzzy setting, De Luca and Termini's non-probabilistic entropy [31] was used to construct fuzzy decision trees [47] and also for feature selection in classification [48]. In both approaches, this entropy is used to select attributes bringing the maximum information.

Another example is the use by Yuan's and Shaw's of the measure of ambiguity issued from Higashi and Klir's measure of ambiguity [39] to build fuzzy decision trees [38].

In these cases, with such measures able to handle fuzzy sets, the R-monotonicity is not the main property needed, but the O-monotonicity is more important when used to compare membership functions.

Shannon entropy is also very frequently used as a regularisation term. It is for instance the case in semi-supervised learning [49]. In this setting, considering a set of variables involved in the optimisation of a given function, entropy regularisation introduces an entropy term in the function to optimise (either maximise or minimise) in order to lead to a sparse distribution on the value (minimisation of the entropy) or a homogeneous distribution (maximisation of

the entropy). Here again, it is the R-monotonicity of the entropy that is called in: the involved process is similar to the one described for the maximum entropy principle.

In unsupervised learning, a popular example of the use of the Shannon entropy lies in a regularisation term for the cost function in fuzzy clustering [50]. This regularisation term should be built on the basis of probabilities, weights, fuzzy memberships,... In this case, there are two possibilities, either the entropic regularisation term must be maximised and leads to a uniform distribution of the related values, or it should be minimised to leads to a distribution with null values in order to introduce sparsity in the trained model.

4.3. Entropies in Other Applications Domains

One notable use of entropies could be found in biology, in the study of ecological systems. The Shannon entropy, as well as the Gini-Simpson index, could be used to evaluate the diversity of the species in an ecosystem [51], [52]. In this framework, each specie is associated with a probability of occurrence that enables the definition of a distribution of probabilities regarding all the species present in the ecosystem. Here, the Shannon entropy applied to this distribution is a suitable tool to evaluate the diversity of the species and enables the comparison among ecosystems or to model species geographic distributions [53], [54].

Similarly to the case of the Maxent principle, it can be highlighted that the R-monotonicity is the property needed in this process.

There are several domains in which the Maxent principle is used. We can cite for instance, non-monotonic reasoning where Shannon entropy is used to build a probability distribution during the decision process in order to select plausible conclusions [55]. Another use can be found in description logics [56] where this principle is also used to make a choice between models.

5. Conclusion

Entropy and measures of information have been extensively studied for 70 years. The original quantities dealing with probabilities of events have been

extended to take into account fuzzy sets, intuitionistic fuzzy sets and other representation models of uncertainty and imprecision. Most of the proposed measures are only based on a formal analogy between the introduced quantities and classic entropies, in spite of the fact that their purpose is different, entropies measuring the decrease of uncertainty resulting from the occurrence of an event, while fuzzy set related measures evaluate the imprecision of events and the fuzziness or non-specificity of the studied observations.

All these quantities have in common a few or many fundamental properties, depending on the case. Various works have listed such properties, for instance [24], [30], [39] and shown which quantities satisfy or do not satisfy them. Attempts have also been done to exhibit classes of quantities with a similar behaviour with regard to sets of properties [57].

In this paper, we highlight the common property of monotonicity of entropy measures with regard to a refinement of information, showing that the main differences between these quantities come from the diversity of orders defining such a refinement. This paper is not intended to provide a review of all entropy measures existing in the literature, but to clarify the concept of refinement of information and the underlying monotonicity, and to illustrate this paradigm by classic examples in a sample of knowledge representation environments, namely the classic probabilistic one, the fuzzy one, the similarity-based one and the intuitionistic fuzzy framework. A focus is put on the importance of monotonicity when entropies or measures of information are used in Artificial Intelligence.

In the future, we will point out new forms of monotonicity useful in Artificial Intelligence and we will provide some hints to choose one or the other measure of information in a given context.

References

[1] C. E. Shannon, The mathematical theory of communication, University of Illinois Press, Urbana, USA, 1948, c. E. Shannon and W. Weaver Eds.

[2] N. Wiener, Cybernetics, or control and communication in the animal and

the machine, 2nd Edition, Hermann & Cie & Camb. Mass. (MIT Press), Paris, 1948.

[3] R. Carnap, Y. Bar-Hillel, An outline of a theory of semantic information, Research Laboratory of Electronics, MIT Technical report NO 247.

[4] E. T. Jaynes, Information theory and statistical mechanics, Physical Review Series II. 106 (4) (1957) 620–630.

[5] S. Kullback, R. Leibler, On information and sufficiency, Annals of Mathematical Statistics 22 (1951) 79–86.

[6] S. Kullback, Information theory and statistics, John Wiley and Sons, NY, 1959.

[7] J. McCarthy, M. L. Minsky, N. Rochester, C. E. Shannon, A proposal for the Dartmouth summer research project on artificial intelligence. August 31, 1955, AI magazine 27 (4) (2006) 12–14.

[8] J. F. Lemmer, S. W. Barth, Efficient minimum information updating for bayesian inferencing in expert systems, in: Proceedings of the AAAI conference, AAAI, 1982, pp. 424–427.

[9] P. Cheeseman, A method of computing generalized bayesian probability values for expert systems, in: Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Vol. 1 of IJCAI'83, Morgan Kaufmann Publishers Inc., 1983, pp. 198–202.

[10] J. R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.

[11] N. J. Nilsson, Probabilistic logic, Artif. Intell. 28 (1) (1986) 71–88. doi:10.1016/0004-3702(86)90031-7.

[12] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[13] M. Goldszmidt, J. Pearl, P. Morris, A maximum entropy approach to non-monotonic reasoning, in: AAAI 1990, AAAI, 1990, pp. 646–652.

[14] R. Jiroušek, P. P. Shenoy, A new definition of entropy of belief functions in the Dempster–Shafer theory, International Journal of Approximate Reasoning 92 (2018) 49–65.

[15] R. Jiroušek, P. P. Shenoy, On properties of a new decomposable entropy of Dempster-Shafer belief functions, International Journal of Approximate Reasoning 119 (2020) 260–279.

[16] G. Zhang, Z. Li, W.-Z. Wu, X. Liu, N. Xie, Information structures and uncertainty measures in a fully fuzzy information system, International Journal of Approximate Reasoning 101 (2018) 119–149.

[17] X.-G. Gao, Z.-G. Guo, H. Ren, Y. Yang, D.-Q. Chen, C.-C. He, Learning bayesian network parameters via minimax algorithm, International Journal of Approximate Reasoning 108 (2019) 62–75.

[18] C. Jiang, D. Guo, Y. Duan, Y. Liu, Strategy selection under entropy measures in movement-based three-way decision, International Journal of Approximate Reasoning 119 (2020) 280–291.

[19] A. E. Allahverdyan, A. Galstyan, A. E. Abbas, Z. R. Struzik, Adaptive decision making via entropy minimization, International Journal of Approximate Reasoning 103 (2018) 270–287.

[20] B. Bouchon-Meunier, C. Marsala, Entropy measures and views of information, in: J. Kacprzyk, D. Filev, G. Beliakov (Eds.), Granular, Soft and Fuzzy Approaches for Intelligent Systems, Vol. 344, Springer, 2017, pp. 47–63.

[21] B. Bouchon-Meunier, C. Marsala, Entropy and monotonicity, in: Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations. IPMU 2018, Communications in Computer and Information Science, vol 854, Springer, 2018, pp. 332–343.

[22] S.-I. Amari, Information geometry in optimization, machine learning and statistical inference, Frontiers of Electrical and Electronic Engineering in China 3 (5) (2010) 241–260.

[23] S. Artstein, K. Ball, F. Barthe, A. Naor, Solution of shannon's problem on the monotonicity of entropy, Journal of the American Mathematical Society 4 (17) (2004) 975–982.

[24] J. Aczél, Z. Daróczy, On Measures of Information and their Characterizations, Vol. 115 of Mathematics in Science and Engineering, Academic Press, New York, 1975.

[25] A. Rényi, On measures of entropy and information, in: Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1: Contributions to the Theory of Statistics, University of California Press, 1961, pp. 547–561.

[26] M. Mugur-Schächter, The general relativity of descriptions, Analyse de Systèmes 11 (4) (1985) 40–82.

[27] J. Kampé de Fériet, Mesures de l'information par un ensemble d'observateurs, in: Gauthier-Villars (Ed.), Comptes Rendus des Scéances de l'Académie des Sciences, Vol. 269 of série A, Paris, 1969, pp. 1081–1085.

[28] J. Kampé de Fériet, Mesure de l'information fournie par un événement, in: Séminaire sur les questionnaires, Publication Structures de l'Information, Université Paris 6, 1971.

[29] J. Aczél, Z. Daróczy, A mixed theory of information. I: Symmetric, recursive and measurable entropies of randomized systems of events, R.A.I.R.O. Informatique théorique / Theoretical Computer Science 12 (2) (1978) 149–155.

[30] G. Klir, M. J. Wierman, Uncertainty-Based Information. Elements of Generalized Information Theory, Studies in Fuzziness and Soft Computing, Springer-Verlag, 1998.

[31] A. de Luca, S. Termini, A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory, Information and Control 20 (1972) 301–312.

[32] S. Guiaşu, Weighted entropy, Reports on Mathematical Physics 2 (3) (1971) 165–179.

[33] L. A. Zadeh, Probability measures of fuzzy events, Journal of Mathematical Analysis and Applications 23 (1968) 421–427.

[34] Z. Daróczy, Generalized information functions, Information and Control 16 (1970) 36–51.

[35] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Taylor and Francis, NY, 1984.

[36] R. R. Yager, Entropy measures under similarity relations, International Journal of General Systems 20 (4) (1992) 341–358.

[37] M. Higashi, G. J. Klir, Measures of uncertainty and information based on possibility distributions, International Journal of General Systems 9 (1) (1982) 43 – 58.

[38] Y. Yuan, M. J. Shaw, Induction of fuzzy decision trees, Fuzzy Sets and Systems 69 (2) (1995) 125 – 139.

[39] G. J. Klir, T. A. Folger, Fuzzy Sets, Uncertainty, and Information, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1987.

[40] P. Rathie, P. Kannappan, A directed-divergence function of type β, Information and Control 20 (1) (1972) 38 – 45.

[41] K. Guo, Q. Song, On the entropy for Atanassov's intuitionistic fuzzy sets: an interpretation from the perspective of amount of knowledge, Applied Soft Computing 24 (2014) 328–340.

[42] E. Szmidt, J. Kacprzyk, New measures of entropy for intuitionistic fuzzy sets, in: Proceedings of the Ninth Int. Conf. on Intuitionistic Fuzzy Sets (NIFS), Vol. 11, Sofia, Bulgaria, 2005, pp. 12–20.

[43] K. T. Atanassov, Intuitionistic fuzzy sets, Fuzzy Sets and Systems 20 (1986) 87–96.

[44] P. Burillo, H. Bustince, Entropy on intuitionistic fuzzy sets and on interval-valued fuzzy sets, Fuzzy Sets and Systems 78 (1996) 305–316.

[45] C. Marsala, B. Bouchon-Meunier, A. Ramer, Ranking attributes to build fuzzy decision trees: a comparative study of measures, in: Proceedings of the eight IFSA'99 World Congress, IFSA, Taipei, Taiwan, 1999, pp. 339–343.

[46] K. Gajowniczek, T. Zabkowski, A. Orłowski, Comparison of decision trees with Rényi and Tsallis entropy applied for imbalanced churn dataset, in: 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE, 2015, pp. 39–44.

[47] K. J. Cios, L. M. Sztandera, Continuous ID3 algorithm with fuzzy entropy measures, in: [1992 Proceedings] IEEE International Conference on Fuzzy Systems, IEEE, 1992, pp. 469–476.

[48] P. Luukka, Feature selection using fuzzy entropy measures with similarity classifier, Expert Systems with Applications 38 (4) (2011) 4600 – 4607.

[49] Y. Grandvalet, Y. Bengio, Semi-supervised learning by entropy minimization, in: Advances in neural information processing systems, 2005, pp. 529–536.

[50] R. Kruse, C. Döring, M.-J. Lesot, Fundamentals of fuzzy clustering, in: J. de Oliveira, W. Pedrycz (Eds.), Advances in Fuzzy Clustering and its Applications, John Wiley and Sons, England, 2007, Ch. 31, pp. 3–30.

[51] R. C. Guiaşu, S. Guiaşu, Entropy in Ecology and Ethology, Nova Science Publishers, Inc., New York, USA, 2003.

[52] R. C. Guiaşu, S. Guiaşu, The weighted Gini-Simpson index: Revitalizing an old index of biodiversity, International Journal of Ecology, 2012, Hindawi Publishing Corporation.

[53] M. P. Austin, Spatial prediction of species distribution: an interface between ecological theory and statistical modelling, Ecological Modelling 157 (2-3) (2002) 101–118.

[54] S. J. Phillips, R. P. Anderson, R. E. Schapire, Maximum entropy modeling of species geographic distributions, Ecological Modelling 190 (3–4) (2006) 231–259.

[55] M. Goldszmidt, P. Morris, J. Pearl, A maximum entropy approach to nonmonotonic reasoning, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (3) (1993) 220–232.

[56] M. Wilhelm, G. Kern-Isberner, A. Ecke, F. Baader, Counting strategies for the probabilistic description logic $\mathcal{ALC}^{\mathsf{me}}$ under the principle of maximum entropy, in: F. Calimeri, N. Leone, M. Manna (Eds.), Logics in Artificial Intelligence, Springer International Publishing, Cham, 2019, pp. 434–449.

[57] B. Bouchon, Entropic models, Cybernetics and Systems: an International Journal 18 (1) (1987) 1–13.