UNCERTAINTY AND INDISTINGUISHABILITY. APPLICATION TO MODELLING WITH WORDS.

Enric Hernández Jiménez.

Under the supervision of Dr. Jordi Recasens.

SUBMITTED TO OBTAIN DE DEGREE OF
DOCTOR IN ARTIFICIAL INTELLIGENCE.

DEPARTAMENT DE LLENGUATGES I SISTEMES INFORMÀTICS
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONA, SPAIN
NOVEMBER 2006

Als meus pares, l'Ana i el Toni.
A la Susanna.

Contents

Chapter 1

Introduction and preliminaries.

It is the nature of all greatness not to be exact.
Edmun Burke.

This dissertation deals with two key concepts: uncertainty and indistinguishability, and with their relationship with the concept of information.

We advocate that uncertainty and indistinguishability are, in fact, head and tail of the same coin so that traditional and new approaches can be developed from both perspectives, interchangeably.

This thesis is a contribution to the study of uncertainty from the point of view of indistinguishability.

1.1 Historical perspective.

Traditionally, western culture has turned around the concept of wisdom understood as the quest for perfect knowledge. Disciplines which could not yield precise and certain knowledge about their matters of concern were left aside as merely speculative, beyond the scope of "educated" inquiry.

Despite of this bias towards perfect knowledge, uncertainty happens to be present at the real world, either at an empirical level as a consequence of resolution limits or lack of reliability of measurement devices, at a cognitive level caused by the vagueness and ambiguity inherent to natural languages, or even at the physical level as the quantum theory has come to prove.

Uncertainty seems to be unavoidable and there is no point in obviating or neglecting its existence. Research efforts should therefore be focused in

its understanding and management.

The emergence of probability theory around the mid-seventeenth century could be considered as the first approach to the formal study of uncertainty.

For almost three hundred years, uncertainty was conceived solely in probabilistic terms. It was around the second half of the twentieth century when things started to change. The appearance of new theories like Fuzzy Set Theory [186] , Evidence Theory [145] or Fuzzy Measure Theory [152] made clear that probability theory could only capture a particular aspect of uncertainty and it was inappropriate in order to conceptualize other facets.

According to Klir, the different types of uncertainty can be classified in the following groups:

- **Uncertainty derived from non specificity**. Is the kind of uncertainty caused by the existence of several interpretations compatible with the available information. The concept of specificity was originally developed within the framework of classical set theory to capture the uncertainty present in situations when it can not be pointed precisely which element, among a given set of elements, is referring the available information.

- **Uncertainty derived from conflict**. This type of uncertainty arises when information is composed of possibly conflicting evidential claims so that they may point in different directions. Measures of conflict usually generalize the Shannon entropy measure [147] by quantifying the expected value of the amount of conflict between these evidential claims.

- **Uncertainty derived from fuzziness**. When vague labels or properties are used (as, for instance, in natural languages), the lack of a sharp boundary between elements fulfilling a given property and elements that do not, originates this new type of uncertainty.

Notwithstanding this classification, the question of the existence of additional types of uncertainty which might not be included in it is still considered an open issue.

1.2 Indistinguishability.

The concept of equality is a fundamental notion in any theory since it is essential to the ability of discerning the objects to whom it concerns, ability which in turn is a requirement for any classification mechanism that might be defined.

Classification, as the process of grouping or clustering according to a certain criterion of similarity, tends to be intimately related to traditional

notions of identity, indiscernibility and indistinguishability. All these concepts have a long tradition as subjects of discussion of fields like philosophy, psychology or even mathematics.

The standard way of approaching the concept of identity is linked to a tradition that can be traced back in time. For instance, Leibniz's Law of Identity is usually written in a second-order language as

$$x \approx y \Leftrightarrow \forall P : P(x) \Leftrightarrow P(y) \tag{1.1}$$

where x and y denote individuals and P ranges over the set of properties.

Leibniz's Law, which is a conjunction of the principles of the Identity of Indiscernible and Indiscernibility of Identical, is intended to express the concept of identity as agreement with respect to all properties. The original postulate may have evolved towards more elaborated formulations based on the idea of the invariance of the set of all automorphisms definable over a given structure, but the main idea behind remains the same.

When all the properties involved are entirely precise (lack of uncertainty), what we obtain is the classical equality, where two individuals are considered equal if and only if they share the same set of properties. What happens, however, when imprecision arises as in the case of properties which are fulfilled only up to a degree? Thus, because certain individuals will be more similar than others, the need for a gradual notion of equality arises.

A further example is when limitations in perceiving and measuring these properties imply the emergence of an approximate equality. Let us, for instance, consider the case of a particular appliance providing measurements on the real line with an error margin ε. It naturally defines the following approximate equality relationship:

$$x \approx y \Leftrightarrow |x - y| \leqslant \varepsilon \tag{1.2}$$

by which two measurements will become distinguishable only if their absolute difference is above the error threshold ε.

Relation $\approx$ is not transitive since we could have $x \approx y$, $y \approx z$ and not necessarily $x \approx z$. In [132], Poincaré was concerned with this apparent paradox. He pointed out that equality satisfies transitivity only in the context of pure mathematics. In the real world, "equal" really means "indistinguishable".

Lack of transitivity also appears when dealing with properties that are inherently vague. Indeed, since a chain of objects that are usually indistinguishable can lead from one which clearly seems to be compatible with a given property to one which clearly does not, the sorites paradox and the corresponding break in transitivity ensue.

It should be remarked the paradigm shift involved, since from Euclid (2300 years ago) transitivity was usually linked to the concept of equality as stated in his "Elements" in the very first "common notion" (self-evident

truth): "Things which are equal to the same thing are also equal to one another."

These considerations show that certain contexts that are pervaded with uncertainty require a more flexible concept of equality that goes beyond the rigidity of the classic concept of equality. Furthermore, since the concept of equivalence relation as the mathematical tool that is used to define the underlying structure is definitely lost, some work around should be provided.

We must deal now with a relation that is reflexive, symmetric and possibly not (in the usual terms) transitive. In this case, the current version of the "triangle inequality" for the underlying metrics translates into a new kind of transitivity property, in which transitivity is defined in terms of a minimum threshold.

Trillas [154] defined an indistinguishability operator in the following manner:

Definition 1.2.1 *An indistinguishability operator on a domain X is defined as a function $E : X \times X \to L$ satisfying $\forall x, y, z \in X$*

1. $E(x, x) \geqslant \lambda$
2. $E(x, y) = E(y, x)$
3. $E(x, y) * E(y, z) \leqslant E(x, z)$

where $(L, \leqslant)$ is a poset, $(L, *)$ is a semigroup and λ is a distinguished element belonging to L.

A special type of indistinguishability operators are obtained particularizing L to be the unit interval with the usual order, and operator $*$ belonging to the class of t-norm functions (see definition (1.5.1))

Definition 1.2.2 *For a given t-norm T, a fuzzy relation E on a set X is a T-indistinguishability operator if and only if for all x, y, z of X the following properties are satisfied:*

1. $E(x, x) = 1$ *(reflexivity)*
2. $E(x, y) = E(y, x)$ *(symmetry)*
3. $E(x, z) \geqslant T(E(x, y), E(y, z))$ *(T-transitivity).*

When the product t-norm is chosen, the resulting class of T-indistinguishability operators equals the class of probabilistic relations studied by Menger [112]. If the Lukasiewicz t-norm is selected, likeness relations as defined by Ruspini [141] are obtained. Finally, similarity relations as introduced by Zadeh [188] are the result of the particularization of T-indistinguishability operators to the minimum t-norm .

T-indistinguishability operators seem to be good candidates for the more flexible and general version of the concept of equality that we are searching for.

1.3 Uncertainty and Information.

The connection between uncertainty and information, as suggested by Klir [95], comes from the fact that uncertainty involved in any problem solving situation is the result of some information deficiency. Therefore, assuming that we can measure the amount of uncertainty associated to a given process, Klir proposes to quantify the information carried by the evidence as the difference between the initial and the final amount of uncertainty, which is measured against the same process when this evidence has been taken into account. In other words, the amount of information can be measured in terms of the reduction of uncertainty.

It should be noted that this characterization based on uncertainty reduction does not capture entirely the richness of the notion of information. However, it has the appealing property of being useful to establish operational criteria to assist in modelling processes.

More precisely, in order to become operational, the following road map is proposed to deal with problems involving some kind of uncertainty:

1.3.1 Modelling.

Dealing with some type of uncertainty requires the existence of a corresponding formal theory that adequately conceptualizes its relevant properties providing, at the same time, a proper language of representation.

Each theory, in turn, implicitly defines a notion of indistinguishability. For instance, the biconditional operator is used in Logic to define equivalence between logical predicates while in geometry, equivalence among figures is based on the notion of congruence, just to mention two examples [154].

Therefore, a first step should be providing a formal framework which properly model the kind of uncertainty we are dealing with, and investigating which notion of indistinguishability is implicitly conveyed by this theory.

1.3.2 Quantification.

The process of building a model accounting for some phenomenon usually involves a balance between accuracy and complexity. The goal is maximazing the accuracy while minimizing the complexity, although both concepts are related in such a way that an increase of accuracy tends to bring about an increase of complexity, and conversely.

Uncertainty also plays a significant role in this trading since a slight increase in the uncertainty associated to the model may often reduce significantly its complexity, at the cost of yielding less precise predictions.

The rules governing the influence of uncertainty in the process of modelling are summarized by the following principles [96]:

- **Principle of minimum uncertainty**. As Yager states, "the more informative way of presenting the data is the data itself" [174]. Any simplification causes a loss of information which results in an increase of uncertainty. This principle prescribes that in case of choosing the model with minimal uncertainty from among a set of candidates, it can be guaranteed the minimal loss of information, thereby maximazing the informational content.

- **Principle of maximum uncertainty**. The principle of minimum uncertainty was intended to govern situations where a simple model compatible with available evidence (represented as a whole set of raw data) is searched. In contrast, the principle of maximum uncertainty is useful when only partial information is available as evidence. Then, commitment to information not entailed by the evidence should be avoided when formulating a model. The principle states that this can be achieved selecting the model which is maximally uncertain with respect to the information not explicitly contained or conveyed by the evidence.[1]

Both principles of minimum and maximum uncertainty rely on the ability of quantifying the amount of uncertainty in order to decide on objective grounds which model has minimum and maximum uncertainty respectively, and should consequently be selected.

1.4 Thesis structure.

The complete program for the study of uncertainty would imply analyzing to which kind of uncertainty each theory is sensitive, and providing definitions and effective procedures for its quantification.

Nowadays, despite of the attainment of remarkable advances, we are still far from the completion of such program. This dissertation should be taken as a contribution towards this goal.

The structure of chapters is in accordance with the road map described in the preceding section.

Chapter 2.

As it was stated previously, each theory implicitly defines its own concept of indistinguishability. The Theory of Evidence arose as a generalization of the theory of probability intended to represent the degree to which available

[1] The well known principle of maximum entropy is just a particularization of the above general principle in the context of classical information theory.

evidence supports the claim that a particular element, whose characterization in terms of the relevant attributes may be deficient, belongs to a given subset of the domain.

This interpretation suggests a notion of indistinguishability between the elements of the domain, depending on their resemblance with that particular element.

In chapter two we will explore further this idea by providing two possible definitions for computing the T-indistinguishability operator associated to a given body of evidence. We will also address the problem of approximating an unrestricted belief measure by a simpler one. Constructive methods for computing the probabilistic and possibilistic approximation will also be introduced. In addition, we will study the dimension of the resulting T-indistinguishability operator and the characterization of an special class of evidences which we have called essentially one-dimensional. Finally, these results will be applied to the field of Game Theory in order to compute the indistinguishability degree between players in a cooperative game.

The main contributions of this chapter have been already published by the author and the advisor of this thesis in the following list of references: [61, 65, 63, 66, 67].

Chapter 3.

From a probabilistic perspective, entropy measures are expected to quantify the uncertainty about the realization of a random variable. In the classical setting, each event is perfectly distinguishable from each other so that each different outcome of this random variable increases its unpredictability.

In chapter three we propose to study the situation in which an indistinguishability relation has been defined over the domain of discourse. Then, the occurrence of two different events but indistinguishable by this indistinguishability relation should count as the occurrence of the same event. Therefore, entropy should be measured with respect distinguishable realizations of this random variable.

We will introduce the observational paradigm and the concept of observational entropy. Definitions of conditional and joint observational entropy will also be provided. Finally, a theorem equivalent to the law of total entropies in the classical setting will be proved.

The contents of this chapter have originated the following publications: [62, 56, 57].

Chapter 4.

The advances in the development of methods for measuring uncertainty have permitted their application in a wide range of areas.

The field of inductive learning has also benefited from the ability of managing uncertainty to extend the range of application domains with those inherently pervaded with some kind of uncertainty, or even to enrich the representation language in order to gain in compactness and expressiveness.

The paradigm of Computing with Words adheres to this approach suggesting the use of words (linguistic labels whose meaning is defined in terms of fuzzy sets) in contrast to the traditional numerical-based computation. However, using linguistic labels require that existing methods should be adapted to cope with them.

From among the family of inductive learning algorithms, decision trees have become one of the most relevant exponents due mainly to their proven applicability to real problems and the readability of the induced knowledge. Generalizations of the basic approach (top-down induction based on information gain heuristics) have also been proposed to deal with uncertainty.

In chapter four we will tackle the task of providing a general framework for induction of decision trees in the presence of uncertainty.

In addition, we will study the application of the concept of observational entropy to the process of building a new type of decision tree: observational decision trees.

Another approach to the induction of rules from data is represented by the family of sequential covering algorithms. The classical approach assumes the use of just crisp information. In this chapter we propose a variant (FSQ) of the general scheme in order to deal with uncertain attributes. Whether the management of uncertainty significatively affects the resulting accuracy will be elucidated by performing a formal comparison over standard data sets.

Finally, we will address the question of how to generate an indistinguishability operator on a domain X which must be compatible with a given indistinguishability operator on set of fuzzy sets (prototypes) over X.

The main contributions of this chapter have been published in the following list of references: [60, 64, 58, 55, 54, 59].

Chapter 5.

Finally, the main contributions of the thesis together with open issues and future lines of research will be summarized in chapter five.

1.5 Preliminaries.

This section provides several definitions and propositions that will be used throughout the dissertation. Some of these results are well known but they are included in order to make this work as self-contained as possible.

1.5.1 On t-norms and T-indistinguishability operators.

Definition 1.5.1 *A function $T : [0,1] \times [0,1] \to [0,1]$ is called a t-norm if the following conditions hold:*

1. $T(x, T(y,z)) = T(T(x,y), z)$ *(associativity)*
2. $T(x,y) = T(y,x)$ *(commutativity)*
3. $x \leqslant x' \Rightarrow T(x,y) \leqslant T(x',y)$
   $y \leqslant y' \Rightarrow T(x,y) \leqslant T(x,y')$
   (monotonicity)
4. $T(x,1) = T(1,x) = x$ *and* $T(0,x) = 0$ *(contour conditions)*

Definition 1.5.2 *A function $S : [0,1] \times [0,1] \to [0,1]$ is called a t-conorm if the following conditions hold:*

1. $S(x, S(y,z)) = S(S(x,y), z)$ *(associativity)*
2. $S(x,y) = S(y,x)$ *(commutativity)*
3. $x \leqslant x' \Rightarrow S(x,y) \leqslant S(x',y)$
   $y \leqslant y' \Rightarrow S(x,y) \leqslant S(x,y')$
   (monotonicity)
4. $S(x,1) = S(1,x) = 1$ *and* $S(0,x) = x$ *(contour conditions)*

Throughout this dissertation, T will denote a continuous t-norm, although most results remain valid assuming only left continuity for T.

Definition 1.5.3 *A t-norm T is Archimedean if and only if the set $\{x \in [0,1] : T(x,x) = x\}$ equals $\{0,1\}$.*

Theorem 1.5.4 *[103] A continuous t-norm T is Archimedean if and only if a strictly decreasing continuous function $t : [0,1] \to [0,+\infty]$ with $t(1) = 0$ exists, such that*

$$T(x,y) = t^{[-1]}(t(x) + t(y)) \tag{1.3}$$

where $t^{[-1]}$ is the pseudo-inverse of t defined as

$$t^{[-1]}(x) = \begin{cases} t^{-1}(x) & \text{if } x \in [0, t(0)] \\ 0 & \text{otherwise.} \end{cases} \tag{1.4}$$

Then function t is called an additive generator of the t-norm T.

Definition 1.5.5 *A t-norm T is strict if the set defined as*

$$NilT = \{x \in (0,1) : \exists m \in N \text{ such that } T^m(x) = 0\} \tag{1.5}$$

equals $\emptyset$, and non-strict if $NilT = (0,1)$. [2]

Definition 1.5.6 *Given a continuous t-norm T, its residuation $\hat{T}$ is defined as:*

$$\forall x, y \in [0,1] : \hat{T}(x|y) = sup\{\alpha \in [0,1] : T(\alpha, x) \leqslant y\}. \tag{1.6}$$

Let us present the residuations for the three most commonly used t-norms.

When $T(x,y) = \min(x,y)$ then

$$\hat{T}(x|y) = \begin{cases} 1 & x \leqslant y \\ y & \text{otherwise.} \end{cases} \tag{1.7}$$

When $T(x,y) = x \cdot y$ then

$$\hat{T}(x|y) = \min(1, \frac{y}{x}). \tag{1.8}$$

When $T(x,y) = \max(x + y - 1, 0)$ then

$$\hat{T}(x|y) = \min(1 - x + y, 1). \tag{1.9}$$

Definition 1.5.7 *Given a continuous t-norm T, its biresiduation $\overleftrightarrow{T}$ is defined as:*

$$\overleftrightarrow{T}(x,y) = \min(\hat{T}(x|y), \hat{T}(y|x)). \tag{1.10}$$

Indistinguishability operators extend the concept of equality to the fuzzy framework and they are also called similarities, fuzzy equalities, fuzzy equivalences, likeness and probabilistic relations depending on the authors and the t-norm used to model transitivity. $E(x,y)$ can be viewed as the degree of similarity or indistinguishability between x and y.

Sometimes reflexivity is not required [72]. In non reflexive indistinguishability operators, $E(x,x)$ is interpreted as the degree of existence of the element x.

[2] T^m is defined by recursion as $T^1(x) = x$ and $T^m(x) = T(T^{m-1}(x), x)$.

Definition 1.5.8 *Given a T-indistinguishability operator E on a set X, a fuzzy subset μ of X is called extensional or observable if and only if for all x, y of X*

$$T(E(x,y),\mu(x)) \leq \mu(y). \tag{1.11}$$

H_E will denote the set of all observable fuzzy subsets with respect to E.

Definition 1.5.9 *Given a T-indistinguishability operator E on a set X and a fuzzy subset μ of X, the fuzzy subset $\phi_E(\mu)$ of X is defined by*

$$\forall x \in X : \phi_E(\mu)(x) = sup_{y \in X} T(E(x,y),\mu(y)). \tag{1.12}$$

Proposition 1.5.10 *[79, 80] The set H_E of observable fuzzy subsets of X with respect to E coincides with the fixed points of ϕ_E (i.e.: $\mu \in H_E$ iff $\phi_E(\mu) = \mu$).*

In [12] the set H_E has been interpreted as the power set of the equivalence classes generated by E and ϕ_E as the natural projection in the sense that any fuzzy subset is sent by ϕ_E to the smallest observable fuzzy subset that contains it.

In fact, one of the most interesting aspects of constructing indistinguishability operators is the possibility of partitioning the universe of discourse X into equivalence classes. At this point, two main directions have been taken for deciding when a set of fuzzy subsets form a partition of X. The first one follows the ideas of Zadeh [188] and defines an equivalence class as the column (also called singleton) of a given T-indistinguishability operator E (the fuzzy subset h_{x_0} defined by fixing an element x_0 in E ($h_{x_0}(x) = E(x_0, x)$) while the second one appears in the early works of Ruspini and Bezdek [6, 142] and forces the elements of the partition (called then a hard-partition) to have no intersection and covering the universe.

Definition 1.5.11 *A set P of fuzzy subsets of a set X is called a partition or coverage of X if and only if there exists an indistinguishability operator E on X such that P is the set of columns of E.*

Definition 1.5.12 *Given a t-norm T and a t-conorm S, a finite set $P = \{\mu_1, \mu_2, ..., \mu_n\}$ of fuzzy subsets of X is a hard-partition of X if and only if*

1. *For all x of X and $i, j = 1, 2, ...n$, $i \neq j$, $T(\mu_i(x), \mu_j(x)) = 0$.*
2. *For all x of X, $S(\mu_1(x), \mu_2(x), ..., \mu_n(x)) = 1$.*

Both definitions generalize the concept of crisp partition in different ways and have their own range of applications. To point out only some of them, definition(1.5.11) has been proved useful in order to give theoretical sound to

the use of triangular and trapezoidal fuzzy numbers [78] while the concept of hard-partition is widely used in fuzzy control [91] and is also the way fuzzy c-means partitions a universe (where the t-norm and t-conorm of Lukasiewicz are tacitly assumed [5]).

In this dissertation, T-indistinguishability operators will play a central role.

One problem commonly faced when studying such relations is how to effectively build them.

The traditional approach relies on computing the transitive closure from a reflexive and symmetric relation. This method, however, has not proved to be fully satisfactory because of the computational cost involved or problems as the "chaining" effect.

These weaknesses were surmounted by the introduction of the Representation Theorem for T-indistinguishability operators .

Theorem 1.5.13 *[158] Let E be a map from $X \times X$ into $[0,1]$ and T be a continuous t-norm . E is a T-indistinguishability operator if and only if a family $\{h_j\}_{j\in J}$ of fuzzy sets exists in X, such that*

$$E(x,y) = \inf_{j\in J} \overleftrightarrow{T}\big(h_j(x), h_j(y)\big). \tag{1.13}$$

Corollary 1.5.14 *[158] Given a fuzzy set μ of X, the fuzzy relation E on X defined for all x, y of X by*

$$E_\mu(x,y) = \overleftrightarrow{T}\,(\mu(x), \mu(y)) \tag{1.14}$$

is a T-indistinguishability operator.

The preceding theorem also allows, in a natural way, the definition of the dimension of a T-indistinguishability operator .

Definition 1.5.15 *A fuzzy set h on X is a generator of a T-indistinguishability operator E if h is an element of any family $\{h_j\}_{j\in J}$ that generates E in the sense of theorem* (1.5.13).

Definition 1.5.16 *Let E be a T-indistinguishability operator . The dimension of E is the minimum of the cardinalities of the generating families of E.*

In the same way, when we are concerned with the notion of order in domains pervaded with uncertainty, a natural generalization of the usual (crisp) preorder relation is given by the concept of T-preorder , which is obtained by simply "dropping" the property of symmetry from the definition of T-indistinguishability operator .

Definition 1.5.17 *[158] A function $P : X \times X \to [0,1]$ is a T-preorder if* $\forall x, y, z \in X$

1. $P(x,x) = 1$
2. $T(P(x,y), P(y,z)) \leqslant P(x,z)$

We also have the corresponding representation theorem and notion of dimension for T-preorder .

Theorem 1.5.18 *[158] Let P be a map from $X \times X$ into $[0,1]$ and T be a continuous t-norm . P is a T-preorder if and only if a family $\{h_j\}_{j\in J}$ of fuzzy sets exists in X, such that:*

$$P(x,y) = \inf_{j\in J} \hat{T}(h_j(x)|h_j(y)) \tag{1.15}$$

Definition 1.5.19 *A fuzzy set h in X is a generator of a T-preorder P if h is an element of any family $\{h_j\}_{j\in J}$ that generates P in the sense of theorem* (1.5.18).

Definition 1.5.20 *Let P be a T-preorder . The dimension of P is the minimum of the cardinalities of the generating families of P in the sense of the previous representation theorem.*

1.5.2 On Possibility Theory.

Definition 1.5.21 *Given a finite set of reference X, a possibility measure Pos is a function $\wp(X) \to [0,1]$ that satisfies*

1. $Pos(\emptyset) = 0$
2. $Pos(X) = 1$
3. *For any family* $\{A_i | A_i \in \wp(X)\} : Pos\big(\bigcup_i A_i\big) = \sup_i Pos(A_i)$.

Definition 1.5.22 *Given a finite set of reference X, a necessity measure Nec is a function $\wp(X) \to [0,1]$ that satisfies*

1. $Nec(\emptyset) = 0$
2. $Nec(X) = 1$
3. *For any family* $\{A_i | A_i \in \wp(X)\} : Nec\big(\bigcap_i A_i\big) = \inf_i Nec(A_i)$.

Given a possibility measure Pos and its dual necessity measure Nec, the following properties hold:

- $Nec(A) = 1 - Pos(\bar{A})$
- $Nec(A) \leqslant Pos(A)$
- $Pos(A \cap B) \leqslant \min[Pos(A), Pos(B)]$
- $Nec(A \cup B) \geqslant \min[Nec(A), Nec(B)]$
- $Pos(A) + Pos(\bar{A}) \geqslant 1$
- $Nec(A) + Nec(\bar{A}) \leqslant 1$
- $\max[Pos(A), Pos(\bar{A})] = 1$
- $\min[Nec(A), Nec(\bar{A})] = 0$

1.5.3 On Demspter-Shafer Theory of Evidence.

The Theory of Evidence [145] arose as a generalization of the Bayesian approach to modeling subjective beliefs and overcame several of its drawbacks, such as the lack of a proper representation of ignorance or the non-symmetric property of the conditioning rule.

Definition 1.5.23 *[145] A function $\wp(X) \rightarrow [0,1]$ is a belief function if and only if it satisfies the following conditions:*

1. $Bel(\emptyset) = 0$.
2. $Bel(X) = 1$.
3. $\forall A_1, \ldots, A_n \subseteq X : Bel(A_1 \cup \ldots \cup A_n) \geqslant \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ I \neq \emptyset}} (-1)^{|I|+1} \cdot Bel(\bigcap_{i \in I} A_i)$

where $|I|$ is the cardinality of I.

Definition 1.5.24 *[145] For a given belief function Bel, its associated plausibility measure Pl is defined as*

$$\forall A \subseteq X : Pl(A) = 1 - Bel(\bar{A}) \tag{1.16}$$

Definition 1.5.25 *[145] A function $m : \wp(X) \rightarrow [0,1]$ is called a basic probability assignment (bpa) whenever*

1. $m(\emptyset) = 0$ *(normalization is assumed).*
2. $\sum_{A \subseteq \wp(X)} m(A) = 1$.

(subsets $A \subseteq X : m(A) > 0$ are called focal elements)

Basic probability assignments, and belief and plausibility measures are related univocally to each other by the following expressions:

$$\forall A \subseteq X : Bel(A) = \sum_{B \subseteq X | B \subseteq A} m(B). \tag{1.17}$$

$$\forall A \subseteq X : Pl(A) = \sum_{B \subseteq X | B \cap A \neq \emptyset} m(B). \tag{1.18}$$

$$\forall A \subseteq X : m(A) = \sum_{B | B \subseteq A} (-1)^{|A-B|} Bel(B). \tag{1.19}$$

Theorem 1.5.26 *[145] Let m be a bpa on $\wp(X)$ and Pl be its associated plausibility measure. Then Pl is a possibility measure if and only if the family of focal elements of $\wp(X)$ is nested.*

Definition 1.5.27 *[190] Every possibility measure Pos on $\wp(X)$ can be uniquely determined by a possibility distribution function $h : X \to [0, 1]$, such that*

$$\forall A \in \wp(X) : Pos(A) = \max_{x \in A} h(x). \tag{1.20}$$

Definition 1.5.28 *Let $m \in M$ be a bpa on X. Then, m is consistent if and only if*

$$\bigcap_{A \subseteq X : m(A) > 0} A \neq \emptyset. \tag{1.21}$$

Chapter 2

Indistinguishability in the framework of Dempster-Shafer Theory of Evidence

This isn't right. This isn't even wrong.
Wolfgang Pauli, on a paper submitted by a physicist colleague.

2.1 Introduction.

In this chapter, we aim at studying the concept of indistinguishability within the framework of the Dempster-Shafer Theory.

The main contributions of this chapter are:

- Definition of the T-indistinguishability operator E_1 associated to a given belief function, based on the concept of covering function.
- A new method for approximating an unrestricted belief function by means of possibilistic (nested focal sets) and probabilistic (singleton focal sets) belief functions. Existence and uniqueness of these approximations are also studied.
- Definition of the natural T-indistinguishability operator E_2 associated to any function $F : \wp(X) \to [0, 1]$. Study of properties and dimensionality of operator E_2 for the particularization of function F to belief

functions.

- Application of previous results to compute indistinguishability degrees between players in a cooperative game.

As it was stated in the introductory chapter and according to Leibniz's law (see expression (1.1)), equality can be conceived as agreement with respect to all properties. This formalization implies the relative nature of the concept of indistinguishability.

Indeed, since every theory defines its own set of relevant properties, the application of Leibniz's law yields different equality criteria depending on the context of discourse so that two elements could be indistinguishable in the framework of a given theory, and perfectly distinguishable in some other theory. Every theory implicitly defines its own notion of equality.

Belief functions, as noted in [96], quantify the degree to which available evidence supports the hypothesis that a particular element of a given domain X, whose characterization in terms of the relevant attributes may be deficient, belongs to a subset of X. This interpretation conveys the notion of indistinguishability, which is definable in terms of the compatibility between the elements of the domain and that particular element.

Moreover, from the representation of evidence by means of its basic probability mass assignment on subsets of X (focal elements), it would be reasonable to explain the distinguishability between two elements a, b by the exclusive support given to any of the two (defining the exclusive support for element a as a function of the masses assigned to the focal elements containing a and not containing b, and analogously for b), while the inclusion of a and b in the same focal element should contribute to an increase in the indistinguishability between them, given that they are indistinguishable for that "portion" of evidence at least.

2.2 A projection-based approximation to the definition of indistinguishability.

In this section, we present a method to calculate the T-indistinguishability operator associated with a given body of evidence based on a particular restriction of the plausibility measure over the set of singletons.

2.2.1 Covering functions.

The general concept of the covering function comes from "projecting" a measure defined in $\wp(X)$ over a subset S of $\wp(X)$. We will have different types of covering functions depending on S and the definition of the projection function.

In this chapter we will use the notion of the commonality number introduced in reference [145] to define a specific type of covering function.

Definition 2.2.1 *[145] Let m be a basic probability assignment in $\wp(X)$. The commonality number $Q : \wp(X) \rightarrow [0,1]$ associated with m is defined as:*

$$\forall A \in \wp(X) : Q_m(A) = \sum_{B \in \wp(X) | A \subseteq B} m(B). \tag{2.1}$$

Quoting Shafer: "$Q(A)$ measures the total probability mass that can move freely to every point of A". Using this definition and following Goodman [46]:

Definition 2.2.2 *[46] Let $\wp_n(X) = \{B : B \in \wp(X) \wedge |B| \leqslant n\}$, $n \geqslant 1$, and m a bpa in $\wp(X)$. Then, the n-point coverage function m_n of m is defined as:*

$$\forall A \in \wp_n(X) : m_n(A) = Q_m(A). \tag{2.2}$$

By varying $1 \leqslant n \leqslant |X|$, we obtain different projections of the measure Q_m. The case $n = 1$ has received especial attention in the literature [46, 29, 86, 143].

Definition 2.2.3 *Let m be a bpa in $\wp(X)$. Its one-point coverage function $\mu_m : X \rightarrow [0,1]$[1] is defined as:*

$$\forall x \in X : \mu_m(x) = Q_m(\{x\}) = \sum_{B \in \wp(X) : x \in B} m(B). \tag{2.3}$$

μ_m is the projection of Q_m over the singleton set, that is, the amount of mass that can be moved to every element x of X.

It is obvious that, because Pl is the plausibility measure associated with m, then:

$$\forall x \in X : \mu_m(x) = Q_m(\{x\}) = Pl(\{x\}). \tag{2.4}$$

The equality above provides an interpretation of μ_m as a fuzzy set in which the membership degree represents the compatibility between a particular element and the evidence, computed as the sum of the masses of the focal elements that are compatible with the element in question.

[1] Also called "consonant projection", "falling shadow", "contour function" or "point-trace".

2.2.2 T-indistinguishability operator E_1.

The idea behind T-indistinguishability operator E_1 is based on using the previously introduced one-point coverage function as an approximation of the original *bpa* in order to generate the intended indistinguishability.

Definition 2.2.4 *Let m be a bpa in X and μ_m its one-point coverage function. The T-indistinguishability operator E_1 is defined as*

$$\forall x, y \in X : E_1(x, y) = \overleftrightarrow{T}(\mu_m(x), \mu_m(y)). \tag{2.5}$$

(E_1 is a T-indistinguishability operator as a trivial consequence of proposition (1.5.14)*)*

Example 2.2.5 *Let m be the bpa in $X = \{a, b, c, d\}$ defined by*

$$\begin{aligned} m(\{a, c\}) &= 0.3 \\ m(\{b, c\}) &= 0.3 \\ m(\{a, b, c\}) &= 0.3 \\ m(\{a, b, d\}) &= 0.1 \end{aligned}$$

Its one-point coverage function μ_m, defined as

$$\forall x \in X : \mu_m(x) = \sum_{A \subseteq X : x \in A} m(A)$$

is

$$\begin{aligned} \mu_m(\{a\}) &= 0.7 \\ \mu_m(\{b\}) &= 0.7 \\ \mu_m(\{c\}) &= 0.9 \\ \mu_m(\{d\}) &= 0.1 \end{aligned}$$

Finally, E_1 (taking the Lukasiewicz t-norm) is:

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 1 & 1 & 0.8 & 0.4 \\ 1 & 1 & 0.8 & 0.4 \\ 0.8 & 0.8 & 1 & 0.2 \\ 0.4 & 0.4 & 0.2 & 1 \end{pmatrix} \end{array}$$

2.3 Equivalence criteria and the issue of belief function approximation.

This section is devoted to showing how the aforementioned procedure for computing the T-indistinguishability operator E_1 associated with a given body of evidence may be used to introduce a new approach to the problem of belief function approximation.

After a short presentation of several significant references, we show how the concept of T-preorder (generated, as is operator E_1 , from the one-point coverage function) is better suited than T-indistinguishability operators when dealing with the uniqueness of the procedure for computing the approximation.

2.3.1 Previous work.

The issue of approximating a given belief measure has been addressed by several authors. The need for such approximations comes from the high computational cost required to manage such measures.

Indeed, given a frame of discernment X, a mass function can have up to $2^{|X|}-1$ focal elements all of which have to be represented explicitly in order to properly capture and combine the evidence they encode.

The combination of two belief functions Bel_1, Bel_2 defined on X is known to be performed in time proportional to

$$|F_{Bel_1}| \cdot |F_{Bel_2}| \cdot |X| \tag{2.6}$$

where F_{Bel_1} and F_{Bel_2} are the set of focal elements of Bel_1 and Bel_2, respectively. Moreover, the number of focal sets of the resulting belief function grows exponentially with respect to the number of focal elements of Bel_1 and Bel_2.

An approximation of a given belief measure is expected to be simpler and well-suited for computational and explanation purposes. A natural way of simplifying a given measure is to reduce the number of focal elements which can be accomplished in different ways.

A first approach is removing and/or clustering similar or "unimportant" focal elements in order to reduce its number. The summarization method [106] leaves the best valued focal elements unchanged while the numerical values (masses) of the remaining focal elements are accumulated and assigned to their union set.

A similar method called $D1$ approximation leaves also the highest focal sets unchanged but the masses of the remaining focal sets are distributed in a specific way among those "highest" sets.

k-l-x approximation method [153] operates in a similar way by incorporating just the highest focal elements, assigning mass zero to the rest of

subsets and finally normalizing the result in order to guarantee that the total amount of mass adds up to 1.

Another approach is the use of clustering techniques for iteratively grouping "similar" focal elements [51, 26].

Other strategy for reducing the number of focal elements is constraining the evidence to belong to a predefined class having a relatively simple form. Two obvious distribution-based measures have been suggested as candidates because of the limitation they impose on the maximum number of focal sets: possibility and probability measures.

Indeed, on a set X whose cardinality is n we need $2^n - 2$ values to define a belief or a plausibility measure from a *bpa* while $n - 1$ (assuming normalization) values suffice for possibility and probability measures.

Concerning probabilistic approximations, the pignistic approximation of a given basic probability assignment m in the frame of discernment X defined by:

$$\forall x \in X : p(x) = \sum_{A \subseteq X : x \in A} \frac{m(A)}{|A|} \tag{2.7}$$

and the one proposed by Voorbraak [161]

$$\forall x \in X : p(x) = \frac{\sum_{A \subseteq X : x \in A} m(A)}{\sum_{B \subseteq X} m(B) \cdot |B|} \tag{2.8}$$

are worthwhile to mention.

Consonant approximations have been studied in detail by Dubois and Prade [33]. In their paper they provide effective procedures for computing inner and outer consonant approximations based on the concepts of weak and strong inclusion between random sets.

From expression (2.6) it can be inferred that not only the number of focal elements but also the cardinality of the domain itself determines the cost of computation. In this sense Denoeux and Yaghlane [27] suggest a novel method based on a hierarchical clustering approach for reducing the size of the frame of discernment in such a way that the loss of information content is minimized.

Other original approach is the proposal by Haenni and Lehmann [50] based on the new concept of incomplete belief functions [2] and allowing resource-bounded computation in which the user determines in advance the maximal time available to computation.

[2] Belief measures having an associate basic probability assignment that does not sum up to 1

From the point of view of belief function combination, Dempster's rule is known to be P-complete in the number of evidential sources [89]. An optimal algorithm for computing Dempster's rule was introduced by Kennes [88].

In order to overcome this computational limitation, different approximation methods have been proposed. Reference [146] tackled the problem of combining *bpa* whose focal sets are either members or complement of members of a hierarchical hypothesis space (for any two sets, their intersection is either empty or one of the sets in it). Efficient exact combination procedures are proposed for this restricted kind of evidences.

Another alternative are Monte Carlo techniques [165] which estimate exact values of belief using outcomes of randomly generated samples.

It should be noted that approximating a general belief measure by means of a simpler one is not for free: it implies a reduction or loss of information. So, the question turns out to which properties should be preserved, ranging from committing an approximation to preserve the amount of uncertainty (based on the principle of uncertainty invariance stated by Klir [83]) to methods which preserve some coherence principles as "only the probable is possible" (for probability-possibility transformations [35]) or the concept of weak and strong inclusion above mentioned.

We present an approximation method based on a novel concept: the preservation of the T-preorder defined by the compatibility degree between the evidence and the singletons set.

2.3.2 Tackling order and uniqueness concerns.

From the previous section, we may conclude that the problem of approximating a given belief function is reduced to providing a simpler approximation whilst ensuring that certain restrictions are fulfilled like limiting the number of focal elements [153, 3], or more sophisticated methods such as the fulfillment of Klir's uncertainty invariance principle [83], among others.

Selecting an approximation method becomes, in a sense, a case of deciding which of the properties conveyed by the evidence should be maintained.

The approach followed to compute E_1 allows us to define a partition in the set of all *bpa* where each class of equivalence contains all *bpa* generating the same T-indistinguishability operator E_1 , thereby preserving the property of being equivalent with respect to their associated T-indistinguishability operator when restricting evidence to the singleton set.

Therefore, given a *bpa* m, any other *bpa* belonging to its class of equivalence could be considered a candidate for its approximation. Since we are interested in "simple" approximations (simpler, at least, than the original *bpa*), we should search among its class of equivalence for "good" candidates. Distribution-based (possibilistic and probabilistic) representations of evidence stand as the best choice. Besides, uniqueness is encouraged in order to make the selection process deterministic.

Unfortunately, it is not true to say that classes of the quotient set (for the above equivalence relation) do have a unique possibilistic canonical element as shown by the following counter-example:

Example 2.3.1 *Let m be the following bpa :*

$$\begin{aligned} m(\{b\}) &= 0.1 \\ m(\{a,c\}) &= 0.6 \\ m(\{a\}) &= 0.3 \end{aligned}$$

Note that m is neither nested nor even consistent. Let m', m'' be defined as

$$\begin{aligned} m'(\{a\}) &= 0.3 \\ m'(\{a,c\}) &= 0.5 \\ m'(\{a,b,c\}) &= 0.2 \end{aligned}$$

and

$$\begin{aligned} m''(\{b\}) &= 0.5 \\ m''(\{b,c\}) &= 0.3 \\ m''(\{a,b,c\}) &= 0.2 \end{aligned}$$

Since all m, m' and m'' generate the same T-indistinguishability operator E_1 (assuming the Lukasiewicz t-norm) given by:

$$\begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{pmatrix} 1 & 0.2 & 0.7 \\ 0.2 & 1 & 0.5 \\ 0.7 & 0.5 & 1 \end{pmatrix} \end{array}$$

and since m' and m'' are nested, clearly m, m' and m'' belong to the same class of equivalence and both m' and m'' are possibilistic evidences related to m.

Therefore, the uniqueness of the possibilistic canonical representative must be discarded.

Moreover, certain areas of application require not just the relative notion of indistinguishability but the concept of order to be preserved. For instance, dealing with a decision-making problem usually involves ranking the set of different alternatives in order to choose the best one according to

a predefined criterion. In other words, we are interested in obtaining the order implicitly defined by the evidence.

All these reasons lead us to consider the notion of order as the key property to be preserved by any approximation, and T-preorder as the appropriate mathematical instrument for dealing with it. The subsequent section will develop this idea further.

2.3.3 Equivalence criteria.

As we have pointed out, any non-trivial approximation of a measure involves a simplification or loss of information and, at the same time, enables equivalence criteria for different bodies of evidence to be established.

In this section, we take a closer look at this idea and emphasize the fact that approximations should be informative enough to provide an order on the elements of the domain X, according to their compatibility with the evidence.

The one-point coverage function seems adequate for this purpose. As previously stated, this function measures the compatibility of each element with the evidence by means of the definition of a fuzzy set; consequently, its membership function (as any other membership function) enables the definition of a natural preorder ($\leqslant_{\mu_m}$) between the elements of the domain. This preorder can be defined in terms of the membership function and the usual order in the unit interval ($\leqslant_{[0,1]}$) as:

$$\forall x, y \in X : x \leqslant_{\mu_m} y \Leftrightarrow \mu_m(x) \leqslant_{[0,1]} \mu_m(y). \tag{2.9}$$

Any equivalence criterion between *bpa* will be required to at least preserve the preorder above between the elements of the domain, that is, any two equivalent *bpa* m, m' should define the same preorder ($\leqslant_{\mu_m} = \leqslant_{\mu_{m'}}$).

Observing this precept, let us consider the following three criteria.

Let m, m' be two *bpa* in $\wp(X)$ and μ_m, $\mu_{m'}$ their one-point coverage functions respectively. Then:

1. m and m' are equivalent if and only if their one-point coverage functions are equal.

2. m and m' are equivalent if and only if the T-preorders generated by μ_m and $\mu_{m'}$ are equal.

3. m and m' are equivalent if and only if the natural preorders (crisp) $\leqslant_{\mu_m}$ and $\leqslant_{\mu_{m'}}$ defined by μ_m and $\mu_{m'}$ are equal.

In order to formalize these definitions, we introduce the following lemma:

Lemma 2.3.2 *[158] Any fuzzy set μ on X generates a T-preorder P_μ in X in the following form:*

$$P_\mu(x,y) = \hat{T}(\mu(x)|\mu(y)) \tag{2.10}$$

This lemma, together with the definitions regarding the concept of T-preorder presented in the introduction, allow us to formally define the equivalence criteria that we stated previously.

Definition 2.3.3 *Let M be the set of all bpa on $\wp(X)$; $m \in M$ and $m' \in M$ be two bpa ; μ_m and $\mu_{m'}$ be their one-point coverage functions; $(\leqslant_{\mu_m})$ and $(\leqslant_{\mu_{m'}})$ the preorders (crisp) on X defined $\forall x, y \in X$ by:*

$$x \leqslant_{\mu_m} y \Leftrightarrow \mu_m(x) \leqslant_{[0,1]} \mu_m(y)$$
$$x \leqslant_{\mu_{m'}} y \Leftrightarrow \mu_{m'}(x) \leqslant_{[0,1]} \mu_{m'}(y)$$

and P_{μ_m} and $P_{\mu_{m'}}$ the (one-dimensional) T-preorders generated by μ_m and $\mu_{m'}$ respectively.

Then we define $R_1, R_2, R_3 \subseteq M \times M$ as the following equivalence relations:

$$(m,m') \in R_1 \Leftrightarrow \forall x \in X : \mu_m(x) = \mu_{m'}(x) \tag{2.11}$$
$$(m,m') \in R_2 \Leftrightarrow \forall x,y \in X : P_{\mu_m}(x,y) = P_{\mu_{m'}}(x,y) \tag{2.12}$$
$$(m,m') \in R_3 \Leftrightarrow (\leqslant_{\mu_m}) = (\leqslant_{\mu_{m'}}) \tag{2.13}$$

(It is trivial to check that $R_1 \subset R_2 \subset R_3$)

Example 2.3.4 *Let m, m' be two bpa in $X = \{a, b, c\}$ defined by:*

$$m(\{a\}) = 0.2$$
$$m(\{b,c\}) = 0.2$$
$$m(\{a,b,c\}) = 0.6$$

and

$$m'(\{a\}) = 0.1$$
$$m'(\{b\}) = 0.1$$
$$m'(\{c\}) = 0.1$$
$$m'(\{a,b,c\}) = 0.7$$

Their one-point coverage functions are $\mu_m(a) = \mu_m(b) = \mu_m(c) = 0.8$ *and* $\mu_{m'}(a) = \mu_{m'}(b) = \mu_{m'}(c) = 0.8$, *respectively.*

Therefore, it holds that

$$\begin{aligned} \forall x \in X : \mu_m(x) = \mu_{m'}(x) &\Rightarrow (m, m') \in R_1 \\ \forall x, y \in X : P_{\mu_m}(x, y) = P_{\mu_{m'}}(x, y) = 1 &\Rightarrow (m, m') \in R_2 \\ \forall x \in X : \mu_m(x) = \mu_{m'}(x) \Rightarrow (\leqslant_{\mu_m}) = (\leqslant_{\mu_{m'}}) &\Rightarrow (m, m') \in R_3 \end{aligned}$$

Example 2.3.5 *Let* m, m' *two bpa in* $X = \{a, b, c\}$ *defined by:*

$$\begin{aligned} m(\{a, b\}) &= 0.5 \\ m(\{a, c\}) &= 0.25 \\ m(\{a, b, c\}) &= 0.25 \end{aligned}$$

and

$$\begin{aligned} m'(\{a\}) &= 0.7 \\ m'(\{b\}) &= 0.2 \\ m'(\{c\}) &= 0.1 \end{aligned}$$

Their one-point coverage functions are

$$\begin{aligned} \mu_m(a) &= 1 \\ \mu_m(b) &= 0.75 \\ \mu_m(c) &= 0.5 \end{aligned}$$

and

$$\begin{aligned} \mu_{m'}(a) &= 0.7 \\ \mu_{m'}(b) &= 0.2 \\ \mu_{m'}(c) &= 0.1 \end{aligned}$$

Therefore,

$$\begin{aligned} \forall x \in X : \mu_m(x) \neq \mu_{m'}(x) &\Rightarrow (m, m') \notin R_1 \\ P_{\mu_m} \neq P_{\mu_{m'}} &\Rightarrow (m, m') \notin R_2 \\ (\leqslant_{\mu_m}) = (\leqslant_{\mu_{m'}}) &\Rightarrow (m, m') \in R_3. \end{aligned}$$

Proposition 2.3.6 *Let m_{ign} be the bpa on $\wp(X)$ representing total ignorance ($m(X) = 1$), m_{unif} be the bpa whose associated belief measure equals the measure of probability defined by the uniform probability distribution on X. Then*

$$(m_{ign}, m_{unif}) \in R_2, R_3. \tag{2.14}$$

Proof 2.3.7 *Trivial.* □

2.3.4 Canonical elements.

In this section we will focus on relation R_2 which was defined as:

$$\forall m, m' \in M : (m, m') \in R_2 \Leftrightarrow \forall x, y \in X : P_{\mu_m}(x, y) = P_{\mu_{m'}}(x, y) \tag{2.15}$$

where M is the set of *bpa* on $\wp(X)$.

It should be noted that this criterion is useful only in situations in which we are just interested in the order relation of a set of elements, which is given by their compatibility with the evidence.

R_2 is an equivalence relation so that each class of equivalence c of the quotient set M/R_2 contains all *bpa* evidentially equivalent.

For example, let X be a set of suspects who may have committed a crime and M a set of *bpa* representing evidence of guilt or innocence. Then, the classes of the quotient set would group evidences producing the same verdict, based on the ranking of guilty of the set of suspects.

Now, the question of whether a possibilistic canonical element exists for each equivalence class seems quite natural. Theorem (2.3.19) will provide an affirmative answer to this question.

Previous results are needed before a proof can be established.

Definition 2.3.8 *Let P be a T-preorder on a set X. For any $x \in X$, the fuzzy subset h_x defined by $\forall y \in X : h_x(y) = P(x, y)$ is called a column of P.*

Lemma 2.3.9 *If P is a one-dimensional T-preorder on a set X, then for particular t-norms (archimedean and minimum), P can be generated by one of its columns (which is clearly a normalized fuzzy set).*

Theorem 2.3.10 *Let T be a continuous archimedean t-norm , t be a generator of T and μ and ν be fuzzy subsets of X. Then, μ and ν generate the same T-preorder if and only if $\forall x \in X$ the following condition holds:*

$$t\mu(x) = t\nu(x) + k_1 \text{ with } k_1 \geqslant \sup_{x \in X}\{-t\nu(x)\}. \tag{2.16}$$

Moreover, if T is non-strict, then $k_1 \leqslant \inf_{x \in X}\{t0 - t\nu(x)\}$.

Proof 2.3.11 $\Rightarrow$ *Given $x, y \in X$, we can suppose $\mu(x) \geqslant \mu(y)$ (which implies $\nu(x) \geqslant \nu(y)$).*

$$P_\mu(x,y) = \hat{T}(\mu(x)|\mu(y)) = t^{-1}(t\mu(y) - t\mu(x)).$$
$$P_\nu(x,y) = t^{-1}(t\nu(y) - t\nu(x)).$$

where $t^{[-1]}$ is replaced by t^{-1}, because all the values in brackets are between 0 and $t(0)$.

If $P_\mu = P_\nu$, then

$$t\mu(y) - t\mu(x) = t\nu(y) - t\nu(x).$$

Therefore

$$t\mu(x) - t\mu(y) = t\nu(x) - t\nu(y).$$

Let us fix $y_0 \in X$. Then $t\mu(x) = t\nu(x) + t\mu(y_0) - t\nu(y_0) = t\nu(x) + k_1$.

$\Leftarrow$) Trivial. $\square$

Corollary 2.3.12 *Let T be the Lukasiewicz t-norm and μ and ν be fuzzy subsets of X. Then, μ and ν generate the same T-preorder on X if and only if $\forall x \in X$:*

$$\mu(x) = \nu(x) + k \text{ with } \inf_{x \in X}\{1 - \nu(x)\} \geqslant k \geqslant \sup_{x \in X}\{-\nu(x)\}. \tag{2.17}$$

Proof 2.3.13 *With the same notations as those of the previous theorem and taking $t(x) = 1 - x$ as a generator of the t-norm ,*

$$1 - \mu(x) = 1 - \nu(x) + k_1 \text{ with } \sup_{x \in X}\{-1 + \nu(x)\} \leqslant k_1 \leqslant \inf_{x \in X}\{\nu(x)\}$$

and therefore

$$\mu(x) = \nu(x) + k \text{ with } \inf_{x \in X}\{1 - \nu(x)\} \geqslant k \geqslant \sup_{x \in X}\{-\nu(x)\}.$$

$\square$

Corollary 2.3.14 *Let T be the product t-norm , and μ and ν fuzzy subsets on X. Then, μ and ν generate the same T-preorder on X if and only if $\forall x \in X$:*

$$\mu(x) = \frac{\nu(x)}{k} \text{ with } k \geqslant \sup_{x \in X}\{\nu(x)\}. \tag{2.18}$$

Proof 2.3.15 *With the same notations as those of the previous theorem and taking* $t(x) = -\ln(x)$ *as a generator of the t-norm*

$$-\ln(\mu(x)) = -\ln(\nu(x)) + k_1 \text{ with } k_1 \geqslant \sup_{x \in X}\{\ln(\nu(x))\}$$

and

$$\mu(x) = \frac{\nu(x)}{k} \text{ with } k \geqslant \sup_{x \in X}\{\nu(x)\}.$$

□

Proposition 2.3.16 *Let* T *be the minimum t-norm and let* μ *be a fuzzy subset on* X *such that exists an element* $x_M \in X$ *holding* $\forall x \in X : \mu(x_M) \geqslant \mu(x)$. *Let* $Y \subseteq X$ *be the set of elements* x *of* X *with* $\mu(x) = \mu(x_M)$ *and* $s = \sup_{x \in X-Y}\{\mu(x)\}$. *A fuzzy subset* ν *on* X *generates the same* T*-preorder as* μ *in* X *if and only if* $\forall y \in Y$*:*

$$\forall x \in X - Y : \mu(x) = \nu(x) \quad \wedge \quad \nu(y) = t \text{ with } s < t \leqslant 1. \tag{2.19}$$

Proof 2.3.17 *It follows trivially from the fact that*

$$P_\mu(x, y) = \begin{cases} \mu(y) & \text{if } \mu(x) \geqslant \mu(y) \\ 1 & \text{if } \mu(x) \leqslant \mu(y). \end{cases} \qquad \square$$

Proposition 2.3.18 *[33] Let* $m \in M$ *be a bpa on* X *and* μ_m *its one-point coverage function. Then* μ_m *is normalized if and only if* m *is consistent.*

Now we can enunciate the following theorem:

Theorem 2.3.19 *Let* T *be a continuous archimedean t-norm or the minimum t-norm , and let* M *be the set of bpa on* X. *Then* $\forall m \in M$ *a unique* $m' \in M$ *exists, such that* m' *is nested and* $(m, m') \in R_2$.

Proof 2.3.20 *The proof has two parts. The first one proves, in a constructive manner, the existence of* m'*, and the second part deals with uniqueness.*

- **Existence of** m'.

Let μ_m *be the one-point coverage function of* m. *From* μ_m*, we generate the* T*-preorder* P_{μ_m} *following lemma* (2.3.2). P_{μ_m} *is a one-dimensional* T*-preorder . We denote by* $h_{P_{\mu m}}$ *the fuzzy set which corresponds to a generating column of* P_{μ_m}.

By lemma (2.3.9), $h_{P_{\mu_m}}$ *is normalized and, consequently, defines a possibility measure* Pos *(see definition* (1.5.27)*). Let* m' *be the bpa corresponding to the measure* Pos.

Theorem (1.5.26) *ensures that* m' *is nested, and it is trivial to check that its one-point coverage function* $\mu_{m'}$ *equals the possibility distribution of the fuzzy set* $h_{P_{\mu_m}}$.

Finally, since both $h_{P_{\mu_m}}$ *and* μ_m *generate the same* T*-preorder* P_{μ_m}*, we have* $(m, m') \in R_2$

- **Uniqueness of** m'.

Due to the fact that every continuous archimedean t-norm is isomorphic to either the Lukasiewicz t-norm or to the product t-norm , we can restrict ourselves to these two t-norms and the minimum to prove the uniqueness of m'.

Let us suppose $n \in M$ *be a nested bpa such that* $m' \neq n$ *and* $(m', n) \in R_2$. *Nested bpa are a particular case of consistent bpa (since nested* $\Rightarrow$ *consistent). By proposition* (2.3.18) $\mu_{m'}$ *and* μ_n *are both normalized. Besides, due to* $(m', n) \in R_2$, $\mu_{m'}$ *and* μ_n *generate the same* T*-preorder . Then:*

1. **Lukasiewicz** t**-norm** *:*

   If $(m', n) \in R_2$ *then by theorem* (2.3.12)

   $$\forall x \in X : \mu_n(x) = \mu_{m'}(x) + \alpha$$

   with

   $$0 \geqslant \alpha \geqslant sup_{x \in X}\{-\mu_{m'}(x)\}$$

   (a) *case* $\alpha = 0$*: then* $\forall x \in X : \mu_n(x) = \mu'_m(x)$. *Because* μ_n *and* μ_m *are normalized, it follows that* $n = m'$.

   (b) *case* $\alpha < 0$*: then*

   $$\begin{aligned} \forall x \in X : \mu_n(x) = \mu_{m'}(x) + \alpha \Rightarrow \\ \forall x \in X : \mu_n(x) < \mu_{m'}(x) \Rightarrow \\ \mu_n \text{ non-normalized } \Rightarrow \\ \text{Contradiction!} \end{aligned}$$

2. **Product** t**-norm** *:*

   If $(m', n) \in R_2$, *by theorem* (2.3.14)

   $$\mu_n(x) = \frac{\mu_{m'}(x)}{k}$$

with

$$k \geqslant sup_{x \in X}\{\mu_{m'}(x)\}$$

and

$$k \geqslant 1$$

then

(a) *case* $k = 1$*: then* $\forall x \in X : \mu_n(x) = \mu_{m'}(x)$*, and because both are normalized, it follows that* $n = m'$*.*

(b) *case* $k > 1$*: then*

$$\begin{aligned} \forall x \in X : \mu_n(x) < \mu_{m'}(x) \Rightarrow \\ \mu_n \textit{non-normalized} \Rightarrow \\ \textit{Contradiction!} \end{aligned}$$

3. **Minimum t-norm** *:*

If $(m', n) \in R_2$*, by theorem* (2.3.16)

$$\forall x \in X : \mu_n(x) = \begin{cases} \mu_{m'}(x) & \text{if } \mu_{m'}(x) < 1 \\ t & \text{if } \mu_{m'}(x) = 1 \end{cases}$$

with

$$sup_{x \in X : \mu_{m'}(x) < 1}(\{\mu_{m'}(x)\}) < t \leqslant 1$$

Then:

(a) *case* $t = 1$*: Then* $\forall x \in X : \mu_n(x) = \mu_{m'}(x)$ *and, because both are normalized, it follows that* $n = m'$*.*

(b) *case* $t < 1$*: Then*

$$\begin{aligned} \forall x \in X : \mu_n(x) < 1 \Rightarrow \\ \mu_n \textit{non-normalized} \Rightarrow \\ \textit{Contradiction!} \end{aligned}$$

Note that when m is not consistent, μ_m is not normalized (see proposition (2.3.18)). In this case $\mu_{m'}$ corresponds to the "normalized" version of μ_m (the normalization strategy depending on the t-norm).

Taking the Lukasiewicz t-norm , $\mu_{m'}$ equals the normal version of μ_m obtained by the normalization procedure suggested by Klir [96], which consists in incrementing, $\forall x \in X$, the value $\mu_m(x)$ by the amount $1 - \text{height}(\mu_m)$.

For the product t-norm , the resulting normalization method corresponds to dividing by the maximum membership value.

Finally, when taking the minimum t-norm , the normalization method reduces to "raise" the membership degrees of the elements that have maximum membership value ($x \in X$, such that $\forall y \in X : \mu_m(x) \geqslant \mu_m(y)$) up to 1.

These considerations should be taken as theoretical justifications for choosing the appropriate normalization procedure for a given context.

This theorem shows that, for any measure of plausibility (belief), we can find one (and just one) measure of possibility (necessity) which is evidentially equivalent when restricting the impact of evidence to the singletons set. Therefore, any evidence (represented by a *bpa*) can be converted into possibilistic evidence ensuring that their compatibility ordering with the singletons set remains unchanged. Its uniqueness allows us to take it as the canonical element of its class of equivalence.

Once answered the question of the existence and uniqueness of the possibilistic approximation, we proceed to study the probabilistic approximation. For any *bpa* m, we look for a *bpa* $p \in M$ evidentially equivalent to m and assigning only mass to singletons.

In this case, we are able to build this (unique) *bpa* p for the product t-norm . For the minimum and the Lukasiewicz t-norm , we must impose additional conditions in order to ensure its existence.

Theorem 2.3.21 *Let M be the set of bpa on $\wp(X)$, $m \in M$ and T be the product t-norm. Then a unique $p \in M$ exists, such that p is a probability distribution and $(m, p) \in R_2$.*

Proof 2.3.22 *Let $m \in M$ be a bpa on $X = \{x_1, \ldots x_n\}$. By theorem* (2.3.19) *a unique nested $m' \in M$ exists such that $(m, m') \in R_2$.*

Denoting by $\mu_{m'}$ the one-point coverage function of m', we will build a bpa p from $\mu_{m'}$ which will only assign mass to singletons and $(m', p) \in R_2$ also.

Then, since R_2 is an equivalence relation:

$$\left.\begin{array}{l}(m, m') \in R_2 \\ (m', p) \in R_2\end{array}\right\} \Rightarrow (m, p) \in R_2.$$

Let us see how p is built from $\mu_{m'}$. We must look for a fuzzy set μ_p such that:

$$\forall x \in X : \mu_p = \frac{\mu_{m'}(x)}{\alpha} \text{ with } \alpha \geqslant 1$$

and

$$\sum_{x \in X} \mu_p(x) = 1$$

These conditions define the following restrictions:

$$\left.\begin{aligned} \mu_p(x_1) &= \frac{\mu_{m'}(x_1)}{\alpha} \\ &\vdots \\ \mu_p(x_n) &= \frac{\mu_{m'}(x_n)}{\alpha} \\ \mu_p(x_1) + \ldots + \mu_p(x_n) &= 1 \end{aligned}\right\}$$

with the following unique solution:

$$\alpha = \sum_{x \in X} \mu_{m'}(x). \tag{2.20}$$

Let p be the bpa such that:

$$\begin{cases} p(A) = 0 & \forall A \subseteq X : |A| > 1 \\ p(\{x\}) = \mu_p(x) & \forall x \in X. \end{cases}$$

Clearly, the one-point coverage of p equals μ_p and given that $\mu_p = \frac{\mu_{m'}(x)}{\alpha}$ with $\alpha \geqslant 1$, corollary (2.3.14) allows us to conclude that $(p, m') \in R_2$. □

Theorem 2.3.23 *Let M be the set of bpa on $\wp(X)$, $m \in M$ and T be the Lukasiewicz t-norm. Then, a unique $p \in M$ exists such that p is a probability distribution and $(m, p) \in R_2$, if and only if:*

$$\frac{\sum_{x \in X} \mu_{m'}(x) - 1}{|X|} \leqslant \inf_{x \in X} \{\mu_{m'}(x)\} \tag{2.21}$$

where $\mu_{m'}$ is the one-point coverage function of the unique nested $m' \in M$ such that $(m, m') \in R_2$.

Proof 2.3.24 *Following the same reasoning used in the proof of theorem* (2.3.21)*, we look for a fuzzy set* μ_p *such that:*

$$\forall x \in X : \mu_p = \mu_{m'}(x) + \alpha \text{ with } 0 \geqslant \alpha \geqslant \sup_x\{-\mu_{m'}(x)\}$$

and

$$\sum_{x \in X} \mu_p(x) = 1.$$

These two conditions define the following restrictions:

$$\left.\begin{aligned} \mu_p(x_1) &= \mu_{m'}(x_1) + \alpha \\ &\vdots \\ \mu_p(x_n) &= \mu_{m'}(x_n) + \alpha \\ \mu_p(x_1) + \ldots + \mu_p(x_n) &= 1 \end{aligned}\right\}$$

These restrictions have a (unique) solution if and only if:

$$\frac{\sum_{x \in X} \mu_{m'}(x) - 1}{|X|} \leqslant \inf_{x \in X}\{\mu_{m'}(x)\} \tag{2.22}$$

and the solution is

$$\forall x \in X : \mu_p(x) = \mu_{m'}(x) + \frac{1 - \sum_{x \in X} \mu_{m'}(x)}{|X|}. \tag{2.23}$$

Let p *be the bpa such that:*

$$\begin{cases} p(A) = 0 & \forall A \subseteq X : |A| > 1 \\ p(\{x\}) = \mu_p(x) & \forall x \in X. \end{cases}$$

Given that $\mu_p = \mu_{m'}(x) + \alpha$ *with* $0 \geqslant \alpha \geqslant \sup_{x \in X}\{-\mu_{m'}(x)\}$*, by corollary* (2.3.12) *we can conclude that* $(p, m') \in R_2$ $\square$

Theorem 2.3.25 *Let* M *be the set of bpa on* $\wp(X)$*,* $m \in M$ *and* T *be the minimum t-norm. Then a unique* $p \in M$ *exists, such that* p *is a probability distribution and* $(m, p) \in R_2$*, if and only if:*

$$\frac{1 - \sum_{x \in X : \mu_{m'}(x) < 1} \mu_{m'}(x)}{Card(\{x \in X : \mu_{m'}(x) = 1\})} > \max_{x \in X}(\{\mu_{m'}(x) : \mu_{m'}(x) < 1\}) \tag{2.24}$$

where $\mu_{m'}$ *is the one-point coverage function of the unique nested* $m' \in M$ *such that* $(m, m') \in R_2$*.*

Proof 2.3.26 *Following the argument of theorem* (2.3.21) *and* (2.3.23), *let us define* μ_p *as the following fuzzy set:*

$$\forall x \in X : \mu_p(x) = \begin{cases} \mu_{m'}(x) & \mu_{m'}(x) < 1 \\ t & \mu_{m'}(x) = 1 \end{cases}$$

with

$$\sum_{x \in X} \mu_p(x) = 1$$

and

$$\max_{x \in X : \mu_{m'}(x) < 1} \big(\mu_{m'}(x)\big) < t \leqslant 1.$$

These restrictions have a (unique) solution if and only if:

$$\frac{1 - \sum_{x \in X : \mu_{m'}(x) < 1} \mu_{m'}(x)}{Card(\{x \in X : \mu_{m'}(x) = 1\})} > \max_{x \in X} \big(\{\mu_{m'}(x) : \mu_{m'}(x) < 1\}\big) \qquad (2.25)$$

and, in this case, the solution is:

$$\forall x \in X : \mu_p(x) = \begin{cases} \mu_{m'}(x) & \mu_{m'}(x) < 1 \\ \frac{1 - \sum_{x \in X : \mu m'(x) < 1} \mu_{m'}(x)}{Card(\{x \in X : \mu_{m'}(x) = 1\})} & \mu_{m'}(x) = 1. \end{cases} \qquad (2.26)$$

Let p *be the bpa such that*

$$\begin{cases} p(A) = 0 & \forall A \subseteq X : |A| > 1 \\ p(\{x\}) = \mu_p(x) & \forall x \in X \end{cases}$$

Then, the one-point coverage function of p *equals* μ_p *and by theorem* (2.3.16) $(p, m') \in R_2$.[3]

□

Theorem (2.3.23) has a nice geometric interpretation.

In $X = \{a_1, a_2, \ldots, a_n\}$, every fuzzy subset μ and probability distribution p can be identified with the points $\big(\mu(a_1), \mu(a_2), ..., \mu(a_n)\big)$ and $\big(p(a_1), p(a_2), ..., p(a_n)\big)$ of $[0, 1]^n$ respectively.

[3]For the current theorem, as for theorem (2.3.21) and (2.3.23) , uniqueness of the solutions (when they do exist) results from the fact that a set of restrictions with a unique solution is solved.

In theorem (2.3.23), a probability distribution p exists if and only if $\forall i = 1, 2, \ldots, n$

$$\mu_p(a_i) \geqslant 0. \tag{2.27}$$

and

$$\mu_{m'}(a_i) + \alpha \geqslant 0 \tag{2.28}$$

and the sum of all these numbers is equal to 1.

This geometric interpretation leads to the following result:

Theorem 2.3.27 *Let m be a bpa on $p(X)$ and μ_m the one-point coverage function of m. A probabilistic distribution p on X with $(m, p) \in R_2$ exists with respect to the Lukasiewicz t-norm , if and only if μ_m belongs to the polytope of $[0, 1]^n$ defined by:*

$$\sum_{i \neq j} x_i + (1 - n) \cdot x_j \leqslant 1, \forall j = 1, 2, \ldots, n. \tag{2.29}$$

Moreover, the probabilistic distributions p on X lie on the hyperplane

$$x_1 + x_2 + \ldots + x_n = 1. \tag{2.30}$$

For $n = 2$, all classes of R_2 contain a probabilistic distribution p.

It is also worth pointing out that the probability distribution μ_p, which we obtain (wherever possible) from the possibility distribution $\mu_{m'}$, fulfills the well-known consistency criterion

$$\forall x \in X : \mu_p(x) \leqslant \mu_{m'}(x). \tag{2.31}$$

in all cases.

When product t-norm is taken, an interesting link with Voorbraak's probabilistic approximation can be proven:

Proposition 2.3.28 *Let M be the set of bpa on $\wp(X)$, and let $m \in M$. Then for the product t-norm , the probabilistic approximation m computed in theorem* (2.3.21) *equals Voorbraak's [161] approximation of m.*

Proof 2.3.29 *It follows easily if we rewrite Voorbraak's Bayesian constant*

$$\sum_{B \subseteq X} m(B) \cdot |B|$$

in terms of an ordered possibility distribution.[4]

$$\sum_{B \subseteq X} m(B) \cdot |B| = \sum_{i=1}^{n} (r_i - r_{i+1}) \cdot i = \sum_{i=1}^{n} r_i. \tag{2.32}$$

□

2.3.5 An example.

Let $X = \{a, b, c, d\}$ and m be the evidence represented by the following *bpa* :

$$\begin{aligned} m(\{a, b\}) &= 0.5 \\ m(\{c, d\}) &= 0.2 \\ m(\{a, b, c, d\}) &= 0.3 \end{aligned}$$

Taking the product t-norm , let us build the possibilistic and probabilistic approximations of m.

The one-point coverage function μ_m associated to m is defined by the following distribution:

$$\begin{aligned} \mu_m(a) = \mu_m(b) &= 0.8 \\ \mu_m(c) = \mu_m(d) &= 0.5 \end{aligned}$$

which in turn generates the following one-dimensional T-preorder P_{μ_m}

$$\begin{array}{c} \\ \text{a} \\ \text{b} \\ \text{c} \\ \text{d} \end{array} \begin{array}{c} \begin{array}{cccc} \text{a} & \text{b} & \text{c} & \text{d} \end{array} \\ \begin{pmatrix} 1 & 1 & 0.625 & 0.625 \\ 1 & 1 & 0.625 & 0.625 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

[4] If we assume the finite universe $X = \{x_1, x_2, \ldots, x_n\}$ and let $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n$ (where $A_i = \{x_1, x_2, \ldots, x_i\}$) be a complete sequence of nested subsets that contains all the focal elements of a possibility measure, the ordered possibility distribution is defined as $\{r_1, r_2, \ldots, r_n\}$ where $r_i = \sum_{k=i}^{n} m(A_k)$. Nested structure implies that $r_i \geqslant r_{i+1}$. In this formulation, possibility distributions are always ordered and $r_1 = 1$ and $r_{n+1} = 0$.

Theorem (2.3.19) ensures that the possibility distribution $h_{P_{\mu_m}}$ corresponding to a generating column of P_{μ_m} defines an unique nested *bpa* m' such that $(m, m') \in R_2$.

Let $h_{P_{\mu_m}}$ be the possibility distribution given by:

$$h_{P_{\mu_m}}(a) = h_{P_{\mu_m}}(b) = 1$$
$$h_{P_{\mu_m}}(c) = h_{P_{\mu_m}}(d) = 0.625$$

Then, we compute its associated nested *bpa* m':

$$m'(\{a, b\}) = 0.375$$
$$m'(\{a, b, c, d\}) = 0.625$$

which is the unique possibilistic approximation of m, such that $(m, m') \in R_2$.

Theorem (2.3.21) provides us with a constructive method for computing the probabilistic approximation of m from the previously computed possibilistic approximation. Namely, let $\mu_{m'}$ be the one-point coverage function of m' defined by:

$$\mu_{m'}(a) = \mu_{m'}(b) = 1$$
$$\mu_{m'}(c) = \mu_{m'}(d) = 0.625$$

and p the following probability distribution:

$$\forall x \in X : p(x) = \frac{\mu_{m'}(x)}{\sum_{y \in X} \mu_{m'}(y)} \tag{2.33}$$

given by

$$p(a) = p(b) = 0.3076$$
$$p(c) = p(d) = 0.1923$$

and defining the following *bpa* m_p:

$$m_p(\{a\}) = m_p(\{b\}) = 0.3076$$
$$m_p(\{c\}) = m_p(\{d\}) = 0.1923.$$

Theorem (2.3.21) ensures that m_p is the unique probabilistic approximation such that $(m, m_p) \in R_2$.

2.4 T-indistinguishability operator E_2 .

Returning to the problem of defining the T-indistinguishability operator associated to a given *bpa* , we should point out some drawbacks of the previously defined operator E_1 .

Although the operator E_1 satisfies the intuitive requirements as posed in the introduction, it should be noted that it is based on an approximation of the original evidence, namely, the one-point coverage function. Despite the fact that this approximation is the optimal consonant approximation under the weak inclusion criterion for random sets (as shown in [33]), it has the drawback consisting in the loss of information with respect to the original evidence.

It thus makes sense to look for an alternative that preserves, as far as is possible, the information conveyed by the evidence. The following results will lead us to this goal.

Lemma 2.4.1 *[158] For all continuous t-norm T and $\forall x, y, z \in X$ it holds:*

$$\overleftrightarrow{T}(x,z) \geqslant T\big(\overleftrightarrow{T}(x,y), \overleftrightarrow{T}(y,z)\big). \tag{2.34}$$

Theorem 2.4.2 *Let F be a function $\wp(X) \to [0,1]$. Then $\forall a, b \in X$, the relation*

$$E(a,b) = \min_{A \in \wp(X-\{a,b\})} \overleftrightarrow{T}\big(F(\{a\} \cup A), F(\{b\} \cup A)\big) \tag{2.35}$$

is a T-indistinguishability operator .

Proof 2.4.3 *Let us prove that E is reflexive, symmetric and T-transitive.*
a) **Reflexivity.** *$\forall a \in X$ it holds that*

$$\begin{aligned} E(a,a) &= \min_{A \in \wp(X-\{a\})} \overleftrightarrow{T}\big(F(\{a\} \cup A), F(\{a\} \cup A)\big) \\ &= 1. \end{aligned}$$

b) **Symmetry.** *Immediate from the symmetry of the operator $\overleftrightarrow{T}$.*

c) **T-transitivity.** *If $\forall a, b, c \in X$ and $\forall Z \in \wp(X - \{a, c\})$, it holds that*

$$\overleftrightarrow{T}\big(F(\{a\} \cup Z), F(\{c\} \cup Z)\big) \geqslant T\big(E(a,b), E(b,c)\big) \tag{2.36}$$

then

$$E(a,c) = \min_{Z\in\wp(X-\{a,c\})} \overleftrightarrow{T}\big(F(\{a\}\cup Z), F(\{c\}\cup Z)\big)$$
$$\geqslant T(E(a,b), E(b,c))$$

and T-transitivity would be proved.

Let us show that the above inequality (2.36) *is indeed satisfied. Let* $Z \in \wp(X-\{a,c\})$. *Then, we can consider two cases:*

1. $b \notin Z$.

   Then

$$Z \in (\wp(X-\{a,c\}) \cap \wp(X-\{b\})) = \wp(X-\{a,b,c\})$$

   and by lemma (2.4.1)

$$\overleftrightarrow{T}\big(F(\{a\}\cup Z), F(\{c\}\cup Z)\big) \geqslant$$
$$\geqslant T\Big(\overleftrightarrow{T}\big(F(\{a\}\cup Z), F(\{b\}\cup Z)\big), \overleftrightarrow{T}\big(F(\{b\}\cup Z), F(\{c\}\cup Z)\big)\Big)$$

   and since

$$\wp(X-\{a,b,c\}) \subset \wp(X-\{a,b\})$$
$$\wp(X-\{a,b,c\}) \subset \wp(X-\{b,c\})$$

   we have

$$\geqslant T\Big(\min_{V\in\wp(X-\{a,b,c\})} \overleftrightarrow{T}\big(F(\{a\}\cup V), F(\{b\}\cup V)\big),$$
$$\min_{W\in\wp(X-\{a,b,c\})} \overleftrightarrow{T}\big(F(\{b\}\cup W), F(\{c\}\cup W)\big)\Big)$$
$$\geqslant T\Big(\min_{U\in\wp(X-\{a,b\})} \overleftrightarrow{T}\big(F(\{a\}\cup U), F(\{b\}\cup U)\big),$$
$$\min_{Y\in\wp(X-\{b,c\})} \overleftrightarrow{T}\big(F(\{b\}\cup Y), F(\{c\}\cup Y)\big)\Big)$$
$$= T(E(a,b), E(b,c)).$$

2. $b \in Z$.

Then $Z \in \wp(X - \{a, c\})$ *and besides* $Z \notin \wp(X - \{a, b, c\})$ *which entails*

$$\begin{aligned} \{a\} \cup (Z - \{b\}) &\in \wp(X - \{b, c\}) \\ \{c\} \cup (Z - \{b\}) &\in \wp(X - \{a, b\}). \end{aligned}$$

By lemma (2.4.1)*:*

$$\begin{aligned} &\overleftrightarrow{T}\big(F(\{a\} \cup Z), F(\{c\} \cup Z)\big) \geqslant \\ &T\Big(\overleftrightarrow{T}\big(F((\{a\} \cup (Z - \{b\})) \cup \{b\}), F((\{a\} \cup (Z - \{b\})) \cup \{c\})\big), \\ &\overleftrightarrow{T}\big(F((\{c\} \cup (Z - \{b\})) \cup \{a\}), F((\{c\} \cup (Z - \{b\})) \cup \{b\})\big)\Big) \end{aligned} \tag{2.37}$$

since trivially

$$\begin{aligned} \{a\} \cup Z &= \{a\} \cup (Z - \{b\}) \cup \{b\} \\ \{c\} \cup Z &= \{c\} \cup (Z - \{b\}) \cup \{b\} \\ \{a\} \cup (Z - \{b\}) \cup \{c\} &= \{c\} \cup (Z - \{b\}) \cup \{a\}. \end{aligned}$$

Moreover, as

$$(\{a\} \cup (Z - \{b\})) \in \wp(X - \{b, c\})$$

and

$$(\{c\} \cup (Z - \{b\})) \in \wp(X - \{a, b\})$$

the expression (2.37) *holds*

$$\begin{aligned} &\geqslant T\Big(\min_{U \in \wp(X - \{b,c\})} \overleftrightarrow{T}\big(F(U \cup \{b\}), F(U \cup \{c\})\big) \\ &\min_{Y \in \wp(X - \{a,b\})} \overleftrightarrow{T}\big(F(Y \cup \{a\}), F(Y \cup \{b\})\big)\Big) \\ &= T(E(b, c), E(a, b)). \end{aligned}$$

□

Corollary 2.4.4 *Let Bel be a belief function on X. Then, the relation*

$$E_2(a,b) = \min_{A \subset \wp(X-\{a,b\})} \overleftrightarrow{T}\big(Bel(\{a\} \cup A), Bel(\{b\} \cup A)\big) \tag{2.38}$$

is a T-indistinguishability operator .

Whilst the belief assigned to two equal subsets of X is obviously the same, for $A, B \subseteq X$ such that $A \neq B$ (and assuming $Bel(A) \neq Bel(B)$), the difference of belief could be "explained" by any of the differences (elements or combinations of elements belonging to A and not belonging to B, and reciprocally) between A and B.

However, if we make these two sets differ exactly in just a pair of elements, that is, $a, b \in X$ such that $a \in A$ and $a \notin B$, $b \in B$ and $b \notin A$, and $A - \{a\} = B - \{b\} = C$, then the existing difference of belief between A and B (when it does happen) can only be explained by differences involving element $\{a\}$ and $\{b\}$, since the rest of elements (C) are the same.

On the basis of this idea, we have defined the indistinguishability degree for any pair a, b of elements as the minimum of the biresiduation (for a given t-norm) between their degrees of belief when both are accompanied by the same set of elements.

2.5 Which fuzzy measure?

A few remarks should be made regarding the generality of theorem (2.4.2). As it does not place any restrictions on the functions it applies to (any function $\wp(X) \to [0,1]$ is allowed), it admits the particularization to a huge range of functions.

Nevertheless, not all these functions will provide intuitive T-indistinguishability operators since these functions are expected to previously convey a proper semantics (in terms of uncertain characterization) which, in a certain way, is transferred to their associated T-indistinguishability operator .

Fuzzy measures, as introduced by Sugeno [152], provide a general framework for the representation of information about uncertain variables. Formally, a fuzzy measure μ on X is a mapping $\mu : \wp(X) \to [0,1]$, such that:

1. $\mu(X) = 1$
2. $\mu(\varnothing) = 0$
3. $A \subseteq B \Rightarrow \mu(A) \leqslant \mu(B)$.

$\mu(E)$ (where $E \in \wp(X)$) is interpreted as a measure of the "available confidence" that the uncertain value attained by a variable V lies in the

subset E. It seems pertinent, therefore, to restrict the kind of functions accepted by theorem (2.4.2) to the more suitable class of fuzzy measures.

Dempster-Shafer Theory provides a framework within which information about a variable whose value is unknown may be represented. Basic probability assignments can also be viewed as structures providing partial information on a family of fuzzy measures that are compatible with them. Typically only two measures from this family are considered, namely the measures of belief and plausibility.

Yager [180] provides a uniform method for characterizing a family of fuzzy measures compatible with a given *bpa* . Let m be a *bpa* with focal elements B_i , $i = 1 \ldots q$. For each focal element B_i, let ω_i be its "allocation vector" of dimension $|B_i|$ whose component $\omega_i(j)$ satisfies the following two conditions:

$$\omega_i(j) \in [0, 1] \tag{2.39}$$

and

$$\sum_{j=1}^{|B_i|} \omega_i(j) = 1. \tag{2.40}$$

Then, a set function μ defined by

$$\forall E \in \wp(X) : \mu(E) = \sum_{j=1}^{q} m(B_j) \cdot \sum_{i=1}^{|B_j \cap E|} \omega_j(i) \tag{2.41}$$

is a fuzzy measure compatible with m.

As Yager notes, a few especial cases are worth pointing out. If $\forall i$: $\omega_i(1) = 1$ then a plausibility measure is obtained; if $\forall i : \omega_i(|B_i|) = 1$, we obtain a belief measure; and if $\omega_i(j) = \frac{1}{|B_i|}$, the resulting fuzzy measure is the one described in [150].

The considerations outlined above show that, even when we are restricted to a Dempster-Shafer structure, a whole family of compatible fuzzy measures can be defined. In addition, as previously stated, the generality of theorem (2.4.2) trivially admits the particularization of any of these measures, as we did with belief measures in corollary (2.4.4).

Why then should we favor belief measures over any other possible measures? Plausibility measures seem to be an obvious alternative, because they are the counterparts of belief measures and are the other most common fuzzy measures associated with a given *bpa* .

Definition 2.5.1 *Let Pl be a plausibility measure on X. We define the T-indistinguishability operator E_3 as*

$$\forall a, b \in X : E_3(a,b) = \min_{A \subset \wp(X-\{a,b\})} \overset{\leftrightarrow}{T}\big(Pl(\{a\} \cup A), Pl(\{b\} \cup A)\big). \quad (2.42)$$

(E_3 is a T-indistinguishability operator as a trivial corollary of theorem (2.4.2)*)*

We conclude this section with a set of results which clarify the relationships between T-indistinguishability operators E_1 , E_2 and E_3 .

Proposition 2.5.2 *Let m be a probabilistic bpa over X. It holds that*

$$\forall a, b \in X : E_1(a,b) = E_2(a,b) = E_3(a,b). \quad (2.43)$$

Proof 2.5.3 *Trivial.* □

The preceding proposition accounts for the probabilistic case. Let us now analyze the case of possibilistic (nested) *bpa* .

Proposition 2.5.4 *Let m be a possibilistic (nested) bpa on X and let μ_m be its one-point coverage function as defined in* (2.2.3)*. Then for all $a, b \in X$ it holds that*

$$E_2(a,b) = \overset{\leftrightarrow}{T}\big(1 - \mu_m(a), 1 - \mu_m(b)\big) \quad (2.44)$$

$$E_3(a,b) = \overset{\leftrightarrow}{T}\big(\mu_m(a), \mu_m(b)\big) = E_1(a,b). \quad (2.45)$$

This proposition shows how, in the nested case, both E_2 and E_3 operators can be defined in terms of the T-indistinguishability operator generated by the one-point coverage fuzzy set (or its complement in the case of E_2). This result naturally matches our expectations provided that, when nested, the possibility measure linked to the *bpa* relates biunivocally to a normal fuzzy set (its associated possibility distribution), so that the resulting indistinguishability is expected to agree with the indistinguishability generated by this fuzzy set.

If we take the Lukasiewicz t-norm we can "refine" the previous result, although we need the following lemma before doing so:

Lemma 2.5.5 *[81] Let T be the Lukasiewicz t-norm and μ and ν be fuzzy sets on X. μ and ν generate the same T-indistinguishability operator if and only if $\forall x \in X$*

$$\mu(x) = \nu(x) + k \qquad with \inf_{x\in X}\{1-\nu(x)\} \geqslant k \geqslant \sup_{x\in X}\{-\nu(x)\} \tag{2.46}$$

or

$$\mu(x) = -\nu(x) + k \qquad with \inf_{x\in X}\{1+\nu(x)\} \geqslant k \geqslant \sup_{x\in X}\{\nu(x)\}. \tag{2.47}$$

Corollary 2.5.6 *Let T be the Lukasiewicz t-norm and m be a bpa on X. Then $\forall a, b \in X$, it holds that*

$$E_2(a,b) = E_3(a,b) \tag{2.48}$$

and by proposition (2.5.4) *if m is nested it holds that*

$$\begin{aligned} E_2(a,b) &= \overleftrightarrow{T}(1-\mu_m(a), 1-\mu_m(b)) \\ &= \overleftrightarrow{T}(\mu_m(a), \mu_m(b)) \\ &= E_3(a,b) \\ &= E_1(a,b) \end{aligned}$$

Finally, let us consider the case of ignorance. It is well known that a disadvantage of probability theory is the lack of a proper representation of ignorance, since its usual representation on the form of uniform distribution entails the acceptance of additional and unjustified assumptions. The Theory of Evidence overcomes this drawback by representing ignorance as the vacuous *bpa* ($m(X) = 1$). It would seem desirable that given the fact that both representations stand when we have no evidence at all that might lead to one element being favored over another, this circumstance gives no clues on how to distinguish between them based on our beliefs. The following proposition formalizes this idea:

Proposition 2.5.7 *Let m be the bpa given by $m(X) = 1$ and p the probabilistic bpa given by the uniform distribution on X*

$$\forall x \in X : p(\{x\}) = \frac{1}{|X|}. \tag{2.49}$$

Then their associated E_1 , E_2 and E_3 operators equal the trivial T-indistinguishability operator defined by

$$\forall x, y \in X : \bar{E^1}(x,y) = 1. \tag{2.50}$$

2.6 Addressing dimensionality.

The representation theorem (1.5.13), in addition to the simplicity of the computations involved (compared to the transitive closure approach), also provides a useful interpretation. If the family of generators are viewed as a set of features or prototypes, the theorem states that a fuzzy relation E is a T-indistinguishability operator if a set of features (whose meaning is formally defined as fuzzy sets on X) "explaining" the distinguishability between the elements in terms of their discrepancy when matching these features, exists. Conversely, from a set of features we can obtain a T-indistinguishability operator that accounts for the degree of indistinguishability between the elements when only these features are taken into account.

Therefore, if we define the dimension as the minimum of the cardinalities of the generating families, it makes sense to study low dimension T-indistinguishability operators since these would allow the necessary computations to be simplified and, what is more important, would afford more clarity to the structure of the operator itself because less features or prototypes would be needed to account for its indistinguishability degrees.

The simplest case occurs when the T-indistinguishability operator can be generated by a single feature (fuzzy set) that conveys all the information needed in such a way that, given any pair of elements, their indistinguishability degrees are defined in terms of their relative compatibility with the generating feature.

A complete set of results of the characterization of one-dimensional T-indistinguishability operators and effective procedures for computing the dimension and minimal families of generators of a given T-indistinguishability operator can be found in [80].

The purpose of this section is to perform a similar study for the T-indistinguishability operator E_2 and provide the necessary and sufficient conditions that a given *bpa* must satisfy in order to generate a one-dimensional E_2 .

Despite the generality of theorem (2.4.2), which allows a broad range of T-indistinguishability operators to be defined on the basis of the t-norm and fuzzy measure involved, from now on we will focus on the T-indistinguishability operator E_2 , thereby assuming the particularization to belief functions and the use of the Lukasiewicz t-norm .

2.6.1 On one-dimensional E_2 operators.

Belief functions are complex mappings and are generally difficult to approximate using simpler and more understandable structures without a significant loss of information.

Nevertheless, *bpa* whose associated E_2 operator is one-dimensional can

be approximated using a single feature that carries exactly the same information, from the point of view of indistinguishability, and summarizes its contents in the form of a mathematical object (fuzzy set) which allows the underlying meaning to be grasped in a more straightforward way. In other words, this fuzzy set may be considered to be the prototype that our distribution of belief is committed to.

Having discussed the motivation behind the study of one-dimensional E_2 operators, the first question that should be elucidated is whether there exist *bpa* generating E_2 operators of more than one dimension in order to prevent their characterization to become a trivial issue. In the case of E_1 T-indistinguishability operators , this characterization makes no sense since all E_1 operators are one-dimensional by definition (they are generated from one-point coverage functions and, consequently, have this function as a generator).

The following example proves the existence of E_2 operators whose dimension are greater than one.

Example 2.6.1 *Let m be the following bpa*

$$\begin{aligned} m(\{a,b,d\}) &= 0.2 \\ m(\{b,c,d\}) &= 0.4 \\ m(\{c,d\}) &= 0.4 \end{aligned}$$

Its associated E_2 operator is

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 1 & 0.6 & 0.4 & 0.2 \\ 0.6 & 1 & 0.6 & 0.6 \\ 0.4 & 0.6 & 1 & 0.8 \\ 0.2 & 0.6 & 0.8 & 1 \end{pmatrix} \end{array}$$

which is not one-dimensional [77].

Since the fact that nested *bpa* generate one-dimensional E_2 operators is a trivial corollary of proposition (2.5.4), a first attempt might involve characterizing one-dimensional E_2 as a certain class of *bpa* satisfying conditions such as nesting or consistency. The following example, together with the previous ones, will help us to discard such an approach.

Example 2.6.2 *Let m be the bpa defined as*

$$m(\{a,b\}) = 0.2$$
$$m(\{c,d\}) = 0.3$$
$$m(\{d\}) = 0.5$$

It generates the following one-dimensional E_2 operator

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 1 & 1 & 0.7 & 0.2 \\ 1 & 1 & 0.7 & 0.2 \\ 0.7 & 0.7 & 1 & 0.5 \\ 0.2 & 0.2 & 0.5 & 1 \end{pmatrix} \end{array}$$

This example proves the existence of one-dimensional E_2 operators whose originating *bpa* is neither nested nor even consistent. This, together with example (2.6.1), refutes the possibility of characterizing one-dimensionality based on criteria like nesting or consistency.

Despite our best efforts, tackling the raw problem of one-dimensionality characterization directly has not proved fruitful. A more manageable approximation that might circumvent this difficulty involves restricting the problem to the characterization of certain, well defined one-dimensional configurations, thereby introducing the concept of essentially one-dimensional configurations (instead of specific *bpa*) that are defined as subsets of the power set of X.

Definition 2.6.3 *Let F be a subset of the power set of X. We consider F to be essentially one-dimensional if and only if E_2 is one-dimensional for all mass assignments that have F as the set of focal sets.*

Example 2.6.4 *Let $X = \{a,b,c\}$. The set $F = \{\{a\},\{c\},\{b,c\}\}$ is not essentially one-dimensional. Consider, for example, the mass assignment*

$$m(\{a\}) = 0.3$$
$$m(\{c\}) = 0.5$$
$$m(\{b,c\}) = 0.2$$

which generates the following non-one-dimensional E_2 operator

$$\begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{pmatrix} 1 & 0.7 & 0.6 \\ 0.7 & 1 & 0.5 \\ 0.6 & 0.5 & 1 \end{pmatrix} \end{array}$$

Nevertheless, configurations that are not essentially one-dimensional can generate a one-dimensional E_2 for particular mass assignments. For instance, in example (2.6.4), consider the mass assignment

$$\begin{aligned} m(\{a\}) &= 0.5 \\ m(\{c\}) &= 0.3 \\ m(\{b,c\}) &= 0.2 \end{aligned}$$

which generates the following one-dimensional E_2 operator

$$\begin{array}{c} \\ \mathrm{a} \\ \mathrm{b} \\ \mathrm{c} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{a} & \mathrm{b} & \mathrm{c} \end{array} \\ \begin{pmatrix} 1 & 0.5 & 0.8 \\ 0.5 & 1 & 0.7 \\ 0.8 & 0.7 & 1 \end{pmatrix} \end{array}$$

Example 2.6.5 *Let $X = \{a, b, c\}$. The set $F = \{\{c\}, \{c, b\}, \{b, a\}\}$ is not essentially one-dimensional. Consider, for example, the mass assignment*

$$\begin{aligned} m(\{c\}) &= 0.4 \\ m(\{c,b\}) &= 0.4 \\ m(\{b,a\}) &= 0.2 \end{aligned}$$

which generates the following non one-dimensional E_2 T-indistinguishability operator

$$\begin{array}{c} \\ a \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{pmatrix} 1 & 0.6 & 0.4 \\ 0.6 & 1 & 0.6 \\ 0.4 & 0.6 & 1 \end{pmatrix} \end{array}$$

Lemma 2.6.6 *Let $a, b \in X$ belong to exactly the same focal sets. Then $E_2(a, b) = 1$.*

Lemma (2.6.8) will prove that, as expected, nested configurations are essentially one-dimensional. First, we need the following lemma regarding one-dimensional T-indistinguishability operators characterization.

Lemma 2.6.7 *[77] A T-indistinguishability operator E is generated by a single function h if and only if there is a total order in $X(\leqslant_*)$ whose first element is a and whose last element is b, such that for any $x, y, z \in X$ with $a \leqslant_* x \leqslant_* y \leqslant_* z <_* b$*

$$T(E(x,y), E(y,z)) = E(x,z) > 0. \tag{2.51}$$

Lemma 2.6.8 *If F is nested, then F is essentially one-dimensional.*

Proof 2.6.9 *Let $A_1 \subset A_2 \subset \cdots \subset A_s$ be the focal sets and $m(A_1), m(A_2), \ldots, m(A_s)$ their respective masses.*

Let $x \in A_i - A_{i-1}$, $y \in A_j - A_{j-1}$, $z \in A_k - A_{k-1}$ with $i \leqslant j \leqslant k$.

$$E_2(x,y) = 1 - \sum_{l=i}^{j-1} m(A_l)$$

$$E_2(y,z) = 1 - \sum_{l=j}^{k-1} m(A_l)$$

$$E_2(x,z) = 1 - \sum_{l=i}^{k-1} m(A_l)$$

Therefore, for the Lukasiewicz t-norm
$T(E_2(x,y), E_2(y,z)) = E_2(x,z)$. □

Lemma 2.6.10 *Let $F = \{A_1, \ldots, A_s\}$ with $A_i \cap A_j = A_k \cap A_l$ for all i, j, k, l with $i \neq j, k \neq l$. Then F is essentially one-dimensional.*

Proof 2.6.11 *Let B be the common intersection of the elements of F* [5]*. If $x_i \in A_i - B$ and $x_j \in A_j - B$, then $E_2(x_i, x_j) = 1 - |m(A_i) - m(A_j)|$.*

If $x_i \in A_i - B$ and $x \in B$, then $E_2(x_i, x) = 1 - \sum_{j \neq i} m(A_j)$.

Let us define the following partial order in X:

If $y \in B$ then $y \geqslant x$ $\forall x \in X$

If $x \in A_i - B$, $y \in A_j - B$, then $x \leqslant y$ if and only if $m(A_i) \leqslant m(A_j)$

If $x \notin A_i$ $\forall i$ then $y \geqslant x$ $\forall y \in X$.

Therefore, if $x \leqslant y \leqslant z$, then $T(E_2(x,y), E_2(y,z)) = E_2(x,z)$.

□

Lemma 2.6.12 *Let $F = \{A_1, \ldots, A_s, B\}$ with $A_i \cap A_j = B$ for all i, j with $i \neq j$. Then F is essentially one-dimensional.*

Proof 2.6.13 *Similar to Lemma* (2.6.10). □

Lemma 2.6.14 *Let $F = \{A_1, \ldots, A_s\}$ with complementary sets of F that satisfy the condition of Lemma* (2.6.10)*. Then F is essentially one-dimensional.*

Lemma 2.6.15 *Let $F = \{A_1, \ldots, A_s, B\}$ with the complementary sets of F that satisfy the condition of Lemma* (2.6.12)*. Then F is essentially one-dimensional.*

[5] If the intersection B is the empty set, then we are in the probabilistic case.

Theorem 2.6.16 *Let F be a subset of the power set of X. F is essentially one-dimensional if and only if F can be split into $F_1, F_2, \ldots, F_s$, the sets of F_i are either nested or satisfy the conditions of one of the Lemmas* (2.6.10), (2.6.12), (2.6.14), (2.6.15) *and the sets of F_i are contained in the sets of F_{i-1} $\forall i = 2, \ldots, s$.*

Proof 2.6.17 $\Leftarrow$ *) Lemmas* (2.6.10), (2.6.12), (2.6.14), (2.6.15)

$\Rightarrow$ *) (Contrareciprocal) If F cannot be split in the way required by the theorem, then either*

1. *$\exists a, c, b \in X$ with $a \in A$, $c \in B$, $b \in C$ with $A, B, C \in F$ and $a \notin B \cup C$, $c \in C - A$ and $b \notin A \cup B$.*

2. *$\exists c, b, a \in X$ with $c \in A$, $c, b \in B$, $b, a \in C$ with $A, B, C \in F$ and $c \notin C$, $b \notin A$ and $a \notin A$.*

In the first case, let $|F|$ denote the cardinality of F. If $|F| = 3$, then Example (2.6.5) *shows that F is not essentially one-dimensional.*

If $|F| \geqslant 4$, let us consider the following mass assignment: $m(A) = 0.3$, $m(B) = 0.39$, $m(C) = 0.3$ and for any other set D of F, $m(D) = \frac{0,01}{|F|-3}$.

Then

$$\begin{aligned} 0.6 &\leqslant E_2(c, b) \leqslant 0.61 \\ 0.69 &\leqslant E_2(c, a) \leqslant 0.7 \\ 0.6 &\leqslant E_2(b, a) \leqslant 0.61 \end{aligned}$$

and therefore E_2 is not one-dimensional.

Second case can be studied in a similar way. $\square$

Corollary 2.6.18 *F is essentially one-dimensional if and only if we cannot find cases* (1) *or* (2) *in F.*

2.7 An application: indistinguishability in Cooperative Games.

The notion of game was introduced as a mathematical abstraction for modelling decision problems in competitive and collaborative situations where each participant has only partial influence over the set of variables governing the final outcome.

The publication of *"Game Theory and Economic Behavior"* by Von Neumann and Morgensten [160] represented the formal inception of Game Theory. Since then it has evolved considerably, allowing its application to a wide set of areas such as economics, politics, psychology, ...

Games are usually divided in cooperative and non-cooperative. Non-cooperative games deal with situations in which players select their optimal strategy based on their guesses about which strategies their opponents are more likely to choose. By contrast, cooperative games promote the bargaining and the formation of coalitions in order to increase the amount of "gain" to be redistributed among its components.

Assuming that there is a transferable gain or utility allowing side payments among the players, the problem is then to agree on how the total amount should be split among them. Such divisions of the total return are expected to be fair and rational, in the sense that the total amount received by the players should equal the maximum amount that could be obtained through collaboration, and no player should receive less than that player could obtain acting alone.

The problem of gain redistribution can be included in the study of the more general notion of interaction between members of a given subset of the set of players.

Let us now recall some basic definitions.

Definition 2.7.1 *Let N, with $|N| \geqslant 2$, be the set of players. Any subset $S \subseteq N$ is called a coalition of players. Sets $\emptyset$ and N are called the empty coalition and the grand coalition, respectively.*

Definition 2.7.2 *A cooperative game is given by the pair (N, v) where N is the set of players and v is the characteristic function of the game given by*

$$v : \wp(N) \longrightarrow \mathbb{R}$$

where $v(S)$ is interpreted as the value (also called worth or power) of coalition S when its members act together as a unit.

Definition 2.7.3 *A game (N, v) is superadditive if and only if for all coalitions $S, T \subseteq N$ such that $S \cap T = \emptyset$ it holds*

$$v(S \cup T) \geqslant v(S) + v(T). \tag{2.52}$$

Definition 2.7.4 *A game (N, v) is monotone if and only if for all coalitions $S, T \subseteq N$:*

$$S \subseteq T \Longrightarrow v(S) \leqslant v(T). \tag{2.53}$$

Definition 2.7.5 *A game (N, v) is simple if and only if for every coalition $S, T \subseteq N$, either $v(S) = 0$ or $v(S) = 1$.*

In a simple game a coalition S is called a winning coalition if $v(S) = 1$ and a losing coalition in other case.

Typical examples of simple games are:

- Majority rule game, where $v(S) = 1$ if and only if $|S| > \frac{n}{2}$, and $v(S) = 0$ otherwise.
- Unanimity game, where $v(S) = 1$ if and only if $S = N$, and $v(S) = 0$ otherwise.
- Dictator game, where given a distinguished player a, $v(S) = 1$ if and only if $a \in S$, and $v(S) = 0$ otherwise.

Definition 2.7.6 *A game (N, v) is symmetric if and only if for all coalition S, $v(S)$ depends only on the number of elements of S, say $v(S) = f(|S|)$ for some function f.*

Definition 2.7.7 *Given a game $G = (N, v)$, we define its associated normal game G' as the game (N, v') with characteristic function v' defined as*

$$\forall S \subseteq N : v'(S) = \frac{v(S)}{\max_{T \subseteq N} v(T)}$$

As noted in [48], the fact that for a player a, the amount of gain or utility received is not (in general) equal to the value $v(\{a\})$ explains why players in N have interest in forming coalitions. For instance, consider another player b and assume that $v(\{a\})$ and $v(\{b\})$ are small whereas $v(\{a, b\})$ is large. Then, a and b have clearly a strong interest in their collaboration.

The concept of interaction tries to capture this idea by taking into account not just their synergic behavior when players a and b are considered alone, but all the coefficients involved when a and b form coalitions with the whole set of players.

Let us present the definition of some interaction indices.

Definition 2.7.8 *The Shapley interaction index of a coalition $S \subseteq N$ in a game $G = (N, v)$ is defined as:*

$$SH(v, S) = \sum_{T \subseteq N - S} \frac{(|N| - |T| - |S|)!|T|!}{(|N| - |S| + 1)!} \cdot \Delta_S v(T)$$

where $\forall S \in N$, $\Delta_S v(T)$ is the discrete S-derivative of v at T which can be defined as

$$\Delta_S v(T) = \sum_{L \subseteq S} (-1)^{(|S| - |L|)} \cdot v(T \cup L).$$

Analogously, the Banzhaf interaction index and the chaining interaction index are defined as

$$B(v,S) = \sum_{T \subseteq N-S} \frac{1}{2^{(|N|-|S|)}} \cdot \Delta_S v(T)$$

and

$$CH(v,S) = \sum_{T \subseteq N-S} s \cdot \frac{(|N|-|S|-|T|)!(|S|+|T|-1)!}{|N|!} \cdot \Delta_S v(T)$$

, respectively.

The Shapley, Banzhaf and Chaining interaction indices are instances of the more general class of probabilistic interaction indices.

Definition 2.7.9 *A probabilistic interaction index of a coalition $S \subseteq N$ in a game $G = (N, v)$ is of the form:*

$$PI(v,S) = \sum_{T \subseteq N-S} p_T^S(N) \cdot \Delta_S v(T)$$

where, for any coalition $S \in N$, the family of coefficients $p_T^S(N)_{T \subseteq N-S}$ forms a probability distribution on 2^{N-S}.

When we compute interaction for coalitions S composed by just one player (one member coalitions), the indices above particularize to the well known Shapley, Banzhaf and chaining values.

Definition 2.7.10 *Given a game $G = (N, v)$, the Shapley value of player $a \in N$ is defined by:*

$$SH(v,a) = \sum_{T \subseteq N-\{a\}} \frac{(|N|-|T|-1)!|T|!}{|N|!} \cdot \Delta_a v(T).$$

The Shapley value [148] is the sole value satisfying the linearity, symmetry, dummy player and efficiency axioms, which makes it particularly suitable to be defined as a fair method for utility redistribution.

The Banzhaf and Chaining values are defined similarly by particularizing the proper definitions to one member coalitions.

2.7.1 Players indistinguishability.

Theorem 2.7.11 *Let $G = (N, v)$ a normal cooperative game. Then for all players $a, b \in N$, the relation*

$$E(a,b) = \min_{A \in \wp(N-\{a,b\})} \overleftrightarrow{T}\big(v(\{a\} \cup A), v(\{b\} \cup A)\big) \tag{2.54}$$

is a T-indistinguishability operator .

The theorem above defines the indistinguishability degree for any pair of players a, b as the minimum of the biresiduation between the values of their respective coalitions when both are accompanied by the same set of players.

As an example, let $G = (\{a, b, c, d\}, v)$ an instance of the dictator game where $v(\{a\}) = v(\{a, b\}) = v(\{a, c\}) = v(\{a, d\}) = v(\{a, b, c\}) = v(\{a, b, d\}) = v(\{a, c, d\}) = v(\{a, b, c, d\}) = 1$ and $v(\{b\}) = v(\{c\}) = v(\{d\}) = v(\{b, c\}) = v(\{b, d\}) = v(\{c, d\}) = v(\{b, c, d\}) = 0$. Then, the T-indistinguishability operator E_2 associated to G is

$$\begin{array}{c} \\ \mathrm{a} \\ \mathrm{b} \\ \mathrm{c} \\ \mathrm{d} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{a} & \mathrm{b} & \mathrm{c} & \mathrm{d} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

which clearly partitions the set of players in two disjoint classes, namely the dictator class ($\{a\}$) and the rest of players ($\{b, c, d\}$).

For symmetric games, given that the value of coalitions do not depend on the particular members composing them but on the cardinal, the following proposition holds:

Proposition 2.7.12 *Let $G = (N, v)$ be a symmetric game. Then its associated operator E_2 equals the indistinguishability operator defined by*

$$\forall x, y \in X : \bar{E^1}(x, y) = 1.$$

As a trivial corollary of this proposition, operator E_2 associated to majority and unanimity games equals indistinguishability operator $\bar{E^1}$.

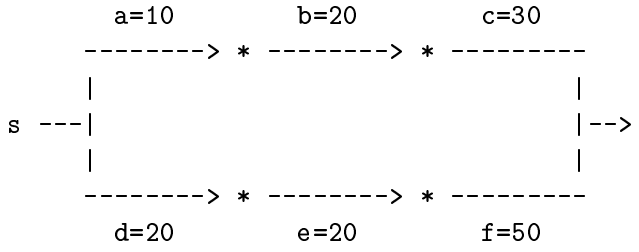
2.7.2 An example: shortest path game.

In this section we will illustrate the presented definitions with the introduction of the shortest path games for which some traffic is supposed to be

routed through a network where each link is owned by a player who incurs some cost while transporting traffic along his link.

The value of a coalition is the payoff paid by sender player (s) when the traffic can be transported (we fix this payoff in 100 units) minus the cost derived from using the links. This cost is supposed to be the minimum, thereby maximizing the net payoff so that if some coalition "opens" two possible paths for routing the traffic, only the cheapest path would be considered when computing the net payoff (i.e value) received by the coalition. Obviously, for coalitions in which sender player is not included, their value is zero since without no sender, no traffic and hence no payoff.

Therefore, the following network



defines the game $G = (\{s, a, b, c, d, e, f\}, v)$ where $\forall S$:

$$v(S) = \begin{cases} 40 & \text{if } \{s, a, b, c\} \in S. \\ 10 & \text{if } \{s, d, e, f\} \in S \text{ and } \{a, b, c\} \notin S. \\ 0 & \text{otherwise.} \end{cases}$$

For game G and taking the Lukasiewicz t-norm, its associated operator E_2 is

$$\begin{array}{c} \\ \text{a} \\ \text{b} \\ \text{c} \\ \text{d} \\ \text{e} \\ \text{f} \\ \text{s} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{a} & \text{b} & \text{c} & \text{d} & \text{e} & \text{f} & \text{s} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0.75 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0.75 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0.75 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0.75 & 0.75 & 0.75 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Note that two classes has naturally formed, namely $\{a, b, c\}$ and $\{d, e, f\}$, corresponding to the two possible paths connecting the origin and destination.

Moreover, sender player s has greater indistinguishability degree with players owning links belonging to the solution path ($s \longrightarrow a \longrightarrow b \longrightarrow c$) than with the rest of players.

2.7.3 Another example: providers game.

In this game, some good is supposed to be produced requiring the integration of several items which, in turn, are offered by their corresponding providers.

For this example, the good to be produced are houses and the set of items assumed to be delivered by the providers a, b, c, d and e are bricks, cement, furniture, paint and fishing equipments, respectively.

As seen at first sight, some items are more important for building houses (brick, cement) than others (furniture, paint), while items such as fishing equipments are completely irrelevant. In fact, without brick and cement there is no house, while paint and furniture are more or less optional, so that providers of basic goods are supposed to be in a better position in eventual negotiations, as much as their expected benefit.

Under this picture, some providers may choose to collaborate in order to maximize their returns.

One possible configuration is the game $G = (\{a, b, c, d, e\}, v)$ where:

$$\begin{aligned} v(\{a, b\}) &= 8 \\ v(\{a, b, c\}) &= 9 \\ v(\{a, b, d\}) &= 9 \\ v(\{a, b, e\}) &= 8 \\ v(\{a, b, c, d\}) &= 10 \\ v(\{a, b, c, e\}) &= 9 \\ v(\{a, b, d, e\}) &= 9 \\ v(\{a, b, c, d, e\}) &= 10 \end{aligned}$$

and for the rest of coalitions $S \in N : v(S) = 0$.

The above game exemplifies the fact that forming coalitions without brick or cement providers are not profitable, while adding optional items like furniture and paint to an already rentable coalition increases in a small degree the value of such coalitions.

Providers of fishing equipments play the role of dummy players since there is no need of such material for building a house and consequently, their joining to a given coalition does not afford additional value at all.

For game G and taking the Lukasiewicz t-norm, its associated operator E_2 is

$$\begin{array}{c} \\ \mathrm{a} \\ \mathrm{b} \\ \mathrm{c} \\ \mathrm{d} \\ \mathrm{e} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathrm{a} & \mathrm{b} & \mathrm{c} & \mathrm{d} & \mathrm{e} \end{array} \\ \begin{pmatrix} 1.0 & 1.0 & 0.1 & 0.1 & 0.0 \\ 1.0 & 1.0 & 0.1 & 0.1 & 0.0 \\ 0.1 & 0.1 & 1.0 & 1.0 & 0.9 \\ 0.1 & 0.1 & 1.0 & 1.0 & 0.9 \\ 0.0 & 0.0 & 0.9 & 0.9 & 1.0 \end{pmatrix} \end{array}$$

Two classes has naturally formed, namely $\{a, b\}$ and $\{c, d\}$, corresponding to the set of providers of basic goods (brick, cement) and the set of providers of optional goods or accessories (furniture, paint).

It can be noted how fishing equipments are more optional than basic in the sense that they are more similar to the class of optional goods than to the class of basic goods. Nevertheless, fishing equipments can not even be considered an optional good (meaning that it provides some additional, although small, value) since they are not entirely indistinguishable from the set of optional goods.

Is interesting to note also that the resulting operator E_2 is one-dimensional.

Given a fixed element e, the "column" fuzzy set μ_e is defined in the following way:

$$\forall p \in N : \mu_e(p) = E_2(e, p).$$

It is trivial to check that the fuzzy set μ_a (column or row corresponding to player a in the matrix representation of E_2) defined by:

$$\begin{aligned} \mu_a(a) &= 1.0 \\ \mu_a(b) &= 1.0 \\ \mu_a(c) &= 0.1 \\ \mu_a(d) &= 0.1 \\ \mu_a(e) &= 0.0 \end{aligned}$$

generates the operator E_2, and consequently E_2 is a one-dimensional indistinguishability operator.

μ_a could be interpreted as the fuzzy set representing the concept "house". The membership degrees of the set of players to μ_a reflect the fact that brick or cement is more essential to the concept "house" than furniture and, obviously, than fish equipments.

The Shapley and Banzhaf values for game G are:

$$SH(v,\{a\}) = 4.66$$
$$SH(v,\{b\}) = 4.66$$
$$SH(v,\{c\}) = 0.33$$
$$SH(v,\{d\}) = 0.33$$
$$SH(v,\{e\}) = 0.0$$

$$B(v,\{a\}) = 4.5$$
$$B(v,\{b\}) = 4.5$$
$$B(v,\{c\}) = 0.25$$
$$B(v,\{d\}) = 0.25$$
$$B(v,\{e\}) = 0.0$$

(For one member coalitions, their Chaining value equals the Shapley value)

Note that values for elements of the same class (basic or optional) are equal which suggests the following proposition:

Proposition 2.7.13 *Given a game $G = (v, N)$ and their associated operator E_2^G, it holds that $\forall a, b \in N$ and for any probabilistic interaction index I:*

$$E_2^G(a,b) = 1 \Rightarrow I(v,\{a\}) = I(v,\{b\}).$$

Chapter 3

Observational Entropy.

Everything is vague to a degree you do not realize till you have tried to make it precise.
Bertrand Russell, "The Philosophy of Logical Atomism".

3.1 Introduction.

Today it is widely assumed that the requirement for informational systems of properly dealing with uncertainty has turned into a must. Hence the necessity of developing methodologies for representing and dealing with uncertainty.

Traditionally, classical set theory and probability theory have been the paradigms of mathematical representations accounting with some aspect of uncertainty. Lately we have witnessed the emergence of a plethora of new theories which have made possible the study of uncertainty from very new points of view.

In this chapter we will summarize significant contributions that can be found in the literature about the study and quantification of uncertainty in the context of different theories.

We will also introduce a novel measure of entropy ("observational entropy") suitable to operate on domains in which an indistinguishability operator has been defined. Properties of this new measure will also be studied.

More specifically, the main contributions of this chapter are:

- Introduction of the "observer paradigm" in order to formalize situations where distinguishability abilities are taken into account when quantifying the predictability of random sources.

- Definition of the concept of observational entropy as a new measure when, besides the usual probabilistic uncertainty, we are dealing with an indistinguishability relation defined over the elements of the domain of discourse.
- Definitions and properties of conditional and joint observational entropy.
- Generalization of the classical law of total entropies to the context of observational entropy.

3.2 Types of uncertainty.

Following the approach of [96] we will consider three types of uncertainty, namely, the one derived from non specificity, conflict and fuzziness. Let us examine them in more detail.

3.2.1 Non specificity.

Non specificity can be seen as the result of the imprecision present in the information, caused by the existence of several interpretations compatible with this information [45].

In other words, available information do not characterize precisely just one designated element but many, thereby inducing more uncertainty as the number of compatible interpretations increases.

The concept of specificity was originally introduced within the framework of classical set theory. As pointed out above, uncertainty derived from non specificity appears when some alternative of particular interest belongs to a whole set of alternatives but we are unable to precise exactly which one of them the available information is referring to.

This considerations led Hartley [52] to quantify the specificity of a given set A as a functional of its cardinal:

$$U(A) = \log_2 |A|. \tag{3.1}$$

Higashi and Klir [69] proposed the measure U-uncertainty as a natural generalization of Hartley measure:

$$U(A) = \int_0^1 \log_2 |A^\alpha| \cdot d\alpha \tag{3.2}$$

where A^α is the α-cut of A, defined as $A^\alpha = \{x \in X | \mu_A(x) \geqslant \alpha\}$.

Yager [170] suggested to define the concept of specificity as a measure of the degree to which a fuzzy set contains one and only one element:

$$U(A) = \int_0^{\alpha_{\max}} \frac{1}{|A^\alpha|} \cdot d\alpha \tag{3.3}$$

where $\alpha_{\max}$ is the largest membership degree in A.

In [176] Yager introduced a family of measures of specificity called linear specificity measures:

$$U(A) = \omega_1 \cdot b_1 - \sum_{j=2}^{n} \omega_j \cdot b_j \tag{3.4}$$

with

$$\begin{aligned} &\omega_j \in [0,1] \\ &\omega_1 = 1 \\ &\omega i \geqslant \omega j \text{ for } i < j \\ &\sum_{j=2}^{n} \omega_j \leqslant 1 \\ &\omega_2 \neq 0. \end{aligned}$$

In this case $U(A)$ is a measure of the degree of satisfaction by a set A of the proposition "A has at least one element and not much more than one" [176]. The weights ω_j are intended to capture the meaning in which the concept "not much more than one" is defined.

A further step in generalization was achieved with the introduction of T-specificity measures [42]:

$$U_T(A) = T_1(a_1, N(T^*_{j=2,\ldots,n}\{T_3(a_j, \omega_j)\})) \tag{3.5}$$

with

$$\begin{aligned} N &= \text{ negation} \\ T_1, T_3 &= t\text{-norms} \\ T^* &= t\text{-conorm} \\ \{\omega_j\} &= \text{ weighting vector} \end{aligned}$$

and $\{a_j\}$ is the ordered vector of membership degrees such that $a_1 \geqslant \ldots \geqslant a_n$.

Those measures represent the notion of "one element (a_1) and no others". In [42] it is proved that both Yager measure of specificity (3.3) and linear

measures of specificity (3.5) are T-specificity measures choosing suitable operators and weights.

Specificity measures based on distances:

$$U(A) = 1 - \min_i(d(A, E_i)) \tag{3.6}$$

where $E_i = (0, \ldots, 1^{(i)}, \ldots, 0)$ and d is a distance, are also particular cases of T-specificity measures.

3.2.2 Conflict.

Uncertainty derived from conflict stems from the generalization of the type of uncertainty measured by Shannon measure.

This statement is based in the fact that if we rewrite Shannon measure

$$U(p) = -\sum_{x \in X} p(x) \cdot \log_2 p(x) \tag{3.7}$$

as

$$U(p) = -\sum_{x \in X} p(x) \cdot \log_2\big(1 - \sum_{y \in X : y \neq x} p(y)\big)$$

the term

$$\sum_{y \in X : y \neq x} p(y)$$

expresses the sum of all evidential claims that fully conflict with the one focusing on x.

The composite expression

$$-\log_2\big(1 - \sum_{y \in X : y \neq x} p(y)\big)$$

just extends the range of conflict quantification from $[0, 1]$ to $[0, \infty)$ supporting the interpretation of Shannon entropy as the expected value of the amount of conflict between evidential claims.

Since measures of conflict have been widely investigated in the framework of the Theory of Evidence, the reader is referred to section 3.3 for a more extensive discussion on this issue.

3.2.3 Fuzziness.

When available evidence is inherently pervaded with vagueness, another kind of uncertainty emerges, namely the uncertainty derived from fuzziness.

In general, a measure of fuzziness is intended to quantify the degree to which the boundary of some fuzzy set is not sharp.

One of the first approaches was introduced by De Luca and Termini [24] who defined the entropy of a fuzzy set A as

$$U(A) = -\sum_{x \in X} \mu_A(x) \cdot \log_2 \mu_A(x) + \big(1 - \mu_A(x)\big) \cdot \log_2 \big(1 - \mu_A(x)\big). \quad (3.8)$$

This measure could be conceived as the aggregation of Shannon entropy of n random variables with just two possible values whose probabilities are $\mu_A(x)$ and $1 - \mu_A(x)$, respectively.

Other approach is the one suggested by Kaufmann [87] consisting in defining the fuzziness of a set A as the distance between its characteristic function and the characteristic function of its "nearest" crisp set (A^{near}) defined by

$$A^{\text{near}} = \begin{cases} 1 & \text{if } \mu_A(x) \geqslant 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (3.9)$$

Then, the measure of fuzziness is defined as

$$U(A) = \frac{2}{|x|^k} \cdot d(A, A^{\text{near}}) \quad (3.10)$$

where d is a metric distance in $[0,1]^X \times [0,1]^X$, and $k \in \mathbb{R}^+$ (k depends on d).

Another way of viewing fuzziness was defined by Yager [169] as the lack of distinction between a set and its complementary: the less the distinction, the fuzzier it is.

$$U(A) = \frac{d(C, \overline{C}) - d(A, \bar{A})}{d(C, \overline{C})} \quad (3.11)$$

where C is an arbitrary crisp set with complement $\overline{C}$. Consequently $d(C, \overline{C})$ is the maximum distance between any pair of sets in $[0,1]^X$.

Kosko [99] associated fuzziness with the ratio of the distances between a given fuzzy set A and its nearest and furthest non fuzzy set:

$$U(A) = \frac{d(A, A^{\text{near}})}{d(A, A^{\text{far}})} \tag{3.12}$$

where A^{near} is the same set as defined in (3.9) and A^{far} its complementary set.

3.2.4 Combined measures.

So far the main concern was to review measures intended to quantify just one kind of uncertainty. Nevertheless, many authors have been worried about developing measures of uncertainty providing an overall aggregate measurement for different types of uncertainty.

One of the first attempts was to define the entropy of a fuzzy set with respect to a discrete probabilistic framework [187]:

$$U(A, p) = -\sum_{x \in X} \mu_A(x) \cdot p(x) \cdot \log_2 p(X) \tag{3.13}$$

which clearly is the Shannon measure weighted by the degrees of membership.

Other line of research was based on the approach of simply trying to aggregate, by means of any suitable algebraic operator, some well established measures of uncertainty of a given kind.

Xie and Bedrosian [167] proposed the sum of Shannon entropy measure (3.7) and De Luca and Termini measure of fuzziness (3.8) as a new combined measure of probabilistic uncertainty and fuzziness:

$$\begin{aligned} U(A, p) = &-\sum_{x \in X} p(x) \cdot \log_2 p(x) \\ &-\sum_{x \in X} \mu_A(x) \cdot \log_2 \mu_A(x) + \big(1 - \mu_A(x)\big) \cdot \log_2\big(1 - \mu_A(x)\big). \end{aligned} \tag{3.14}$$

De Luca and Termini [24] modified the definition above weighting the second term by the appropriate probability measures:

$$\begin{aligned} U(A, p) = &-\sum_{x \in X} p(x) \cdot \log_2 p(x) \\ &-\sum_{x \in X} p(x) \cdot \Big(\mu_A(x) \cdot \log_2 \mu_A(x) + \big(1 - \mu_A(x)\big) \cdot \log_2\big(1 - \mu_A(x)\big)\Big). \end{aligned} \tag{3.15}$$

However, composite measures are meaningful only as long as the "elementary" measures aggregated are additive in nature. This is not always the case since fuzziness is conceptually different from probabilistic uncertainty, making the algebraic combination of expressions (3.14) and (3.15) hard to justify.

More examples of combined measures can be found in section 3.3.4.

3.3 Uncertainty measures in Dempster-Shafer Theory of Evidence.

Dempster-Shafer Theory of Evidence has proved to be a particular fruitful area in the task of investigating uncertainty measures due mainly to its ability to represent and deal with uncertainty of different types.

Indeed, when there is at least one focal set containing more than one element, uncertainty derived from non specificity emerges. Also, if there is more than one focal element, this introduces disagreement in the evidence and, consequently, uncertainty derived from conflict arises, being the conflict larger as the more mutually disagreeing pieces of evidence were.

In this section we will present a review of measures of uncertainty for the Theory of Evidence.

3.3.1 Non specificity measures in the Theory of Evidence.

In section 3.2.1 it was shown how Hartley measure of non specificity was generalized to fuzzy set theory by Higashi and Klir [69].

Dubois and Prade [31] further generalize it to measure non specificity in the Theory of Evidence as follows:

$$U(m) = \sum_{A \subseteq X, A \neq \emptyset} m(A) \cdot \log_2 |A| \tag{3.16}$$

where m is a basic probability assignment.

A modified version was provided by Lamata and Moral [101]:

$$U(m) = \log_2 \Big(\sum_{A \subseteq X} m(A) \cdot \frac{1}{|A|} \Big). \tag{3.17}$$

Yager measure of specificity has also its counterpart in the Theory of Evidence:

$$U(m) = \sum_{A \subseteq X, A \neq \emptyset} m(A) \cdot \frac{1}{|A|}. \tag{3.18}$$

It can be noted that all the presented definitions are a weighted average of some classical measure over the focal elements.

3.3.2 Measures of conflict in the Theory of Evidence.

Unlike non specificity measures (where expression (3.16) is well established as non specificity measure in the Theory of Evidence since uniqueness has been proved on axiomatic grounds), the search for a counterpart of Shannon-like measure is still an open problem.

Although many definitions have been proposed, each has been found to violate some essential requirement. In this section we will summarize some of these definitions.

As shown in section 3.2.2, Shannon measure can be rewritten in terms of averaging conflict between evidential claims as follows:

$$U(m) = \sum_{A \subseteq X} m(A) \cdot \log(\text{ amount of mass consistent with } A). \tag{3.19}$$

Different measures can thus be defined depending on which consistency criterion is considered.

The more restrictive consistency criterion is expressed by

$$\forall A \subseteq X : B \subseteq X \text{ is consistent with } A \Leftrightarrow B \subseteq A$$

Adapting expression (3.19) to this definition of conflict, Höhle measure of confusion [71] is obtained:

$$U(m) = -\sum_{A \subseteq X} m(A) \cdot \log Bel(A). \tag{3.20}$$

A laxer definition of conflict is expressed by

$$\forall A \subseteq X : B \subseteq X \text{ is consistent with } A \Leftrightarrow A \cap B \neq \emptyset$$

This results in Yager measure of dissonance [171]:

$$U(m) = -\sum_{A \subseteq X} m(A) \cdot \log Pl(A). \tag{3.21}$$

Some intermediate criteria have also been proposed defining consistency between B and A as the degree of inclusion of B in A:

$$\text{Consistency}(A,B) \ = \frac{|A \cap B|}{|B|}. \tag{3.22}$$

In this case, Klir and Ramer measure of discordance [94] is obtained :

$$U(m) = \sum_{A \subseteq X} m(A) \cdot \log \sum_{B \subseteq X} m(B) \cdot \frac{|A \cap B|}{|B|}. \tag{3.23}$$

On the contrary, if consistency between A and B is defined as

$$\text{Consistency}(A,B) \ = \frac{|A \cap B|}{|A|} \tag{3.24}$$

then the measure of strife [97] is obtained:

$$U(m) = \sum_{A \subseteq X} m(A) \cdot \log \sum_{B \subseteq X} m(B) \cdot \frac{|A \cap B|}{|A|}. \tag{3.25}$$

Some other measures are also worthwhile to mention like the measure resulting of substituting probabilities by masses in Shannon measure [122, 127]:

$$U(m) = -\sum_{A \subseteq X} m(A) \cdot \log m(A) \tag{3.26}$$

or the measure proposed by Smets [149] based on the commonality number [145]:

$$U(m) = -\sum_{A \subseteq X} \log Q(A). \tag{3.27}$$

3.3.3 Measures of fuzziness in the Theory of Evidence.

Strictly speaking, measures of fuzziness should only be defined for fuzzy sets. Nevertheless, in this section we present a measure of fuzziness for the Theory of Evidence proposed by Dubois and Prade [32] based on Yager's view of fuzziness as lack of distinction between a set and its complementary.

Taking Shafer definitions of complement and conflict between basic probability assignments [145] and defining fuzziness associated to a given *bpa* m as the degree of conflict between m and its complementary, Dubois and Prade suggest the following definition:

$$U(m) = -\ln\Big(\sum_{A,B \subseteq X : A \subseteq B} m(A) \cdot m(B)\Big). \tag{3.28}$$

3.3.4 Combined measures in the Theory of Evidence.

As it was previously stated, some remarkable attempts have been carried out to provide measures of uncertainty capturing in an uniform way different types of uncertainty. Those attempts range from straightforward algebraic combinations of well defined measures of uncertainty to more elaborated definition involving optimization problems.

Representative of the former approach is the proposal by Klir and Ramer [97] consisting in the aggregation of Yager measure of specificity (3.18) and the measure of strife (3.25) which yields the following expression:

$$U(m) = \sum_{A \subseteq X} m(A) \cdot \log \frac{|A|^2}{\sum_{B \subseteq X} m(B) \cdot |A \cap B|}. \tag{3.29}$$

Lamata and Moral [101] propose the aggregation of Yager measure of specificity (3.18) and the measure of dissonance (3.21):

$$U(m) = -\sum_{A \subseteq X} m(A) \cdot \log \frac{|A|}{Pl(A)}. \tag{3.30}$$

Yet another example of algebraic combination is the aggregation of Pal measure of entropy (3.26) and Yager measure of specificity (3.18) proposed by Hemasinha et al [127]:

$$U(m) = \sum_{A \subseteq X} m(A) \cdot \log \frac{|A|}{m(A)}. \tag{3.31}$$

These kind of measures usually exhibit unappealing features since they aggregate measures which are not necessarily additive in nature.

A measure (AU) of total uncertainty [1] ,overcoming some deficiencies present in other candidates, was defined in terms of the solution to a non linear optimization problem.

AU is defined for each belief function Bel as

$$AU(Bel) = \max_{P_{Bel}} \Big[-\sum_{x \in X} p(x) \cdot \log p(x) \Big] \tag{3.32}$$

where P_{Bel} is the set of probability distributions satisfying the consistency criterion expressed by

[1] The expression "total uncertainty" is used in the literature to designate measures which capture more than one type of uncertainty in spite of they may not be an indicator of really the total uncertainty.

$$\forall A \subseteq X : Bel(A) \leqslant \sum_{x \in A} p(x) \tag{3.33}$$

Among other advantages it satisfies the commonly required properties (including subadditivity) for a measure of this type. In addition, the existence and correctness of an algorithmic procedure for its computation have been also proved [96].

Although these desirable properties, measure AU has the shortcoming of being highly insensitive to changes in the evidence as shown in [95].

In order to overcome this drawback, Smith [151] suggests three alternatives to measure total uncertainty. Let $\overline{S}$ and $\underline{S}$ be the maximum and minimum Shannon measure within all probability distributions consistent with the given evidence (assuming the consistence criterion expressed by (3.33)) and N be Dubois and Prade measure of non specificity (3.16).

The first measure U_1 is a linear combination of $\overline{S}$ and N:

$$U_1(m) = \delta \cdot \overline{S_m} + (1 - \delta) \cdot N(m) \tag{3.34}$$

where $\delta \in (0, 1)$.

U_2 is defined as the pair

$$U_2(m) = \big(N(m), \overline{S_m} - N(m)\big). \tag{3.35}$$

U_3 is the pair

$$U_3(m) = \Big(N(m), \big[\overline{S_m} - \underline{S_m}\big]\Big) \tag{3.36}$$

where the second component is the whole range of values of m-consistent Shannon measures.

Up to now all these measures still have aspects that need to be further clarified in order to be meaningful. The quest of the search for a fully satisfactory measure of total uncertainty for Dempster-Shafer Theory of Evidence is rather far from being a closed problem.

3.3.5 Particularization to Possibility Theory.

The Theory of Evidence includes, as a particular case, the Possibility Theory. It is also known that every possibility measure over a finite domain X is univocally determined by a possibility distribution.

In the following table we rewrite some previously presented measures of uncertainty for the Theory of Evidence in terms of a normal possibility distribution r represented by the ordered vector $r = (r_1, \ldots, r_n)$ such that $1 = r_1 \geqslant \ldots \geqslant r_n$.

Measure	**Expression**	**Type**
U-uncertainty (3.16)	$\sum_{i=1}^{n}(r_i - r_{i+1}) \cdot \log i$	non specificity
Measure (3.18)	$\sum_{i=1}^{n}(r_i - r_{i+1}) \cdot \frac{1}{i}$	non specificity
Confusion (3.20)	$-\sum_{i=1}^{n}(r_i - r_{i+1}) \cdot \log(1 - r_{i+1})$	conflict
Dissonance (3.21)	0 (nesting *bpa*)	conflict
Strife (3.25)	$-\sum_{i=2}^{n}(r_i - r_{i+1}) \cdot \log \frac{i}{\sum_{j=1}^{i} r_j}$	conflict
Discordance (3.23)	$\sum_{i=1}^{n-1}(r_i - r_{i+1}) \cdot \log\left[1 - i \cdot \sum_{j=i+1}^{n} \frac{r_j}{j \cdot (j-1)}\right]$	conflict
Measure (3.26)	$-\sum_{i=1}^{n}(r_i - r_{i+1}) \cdot \log(r_i - r_{i+1})$	conflict
Measure (3.29)	$\sum_{i=2}^{n}(r_i - r_{i+1}) \cdot \log \frac{i^2}{\sum_{j=1}^{i} r_j}$	combined
Measure (3.31)	$\sum_{i=1}^{n}(r_i - r_{i+1}) \cdot \log \frac{i}{(r_i - r_{i+1})}$	combined

3.3.6 Particularization to Probability Theory.

When focal elements are singletons, both measures *Bel* and *Pl* collapse in the same probability measure.

Another consequence is the lack of uncertainty derived from non specificity since all focal elements are singletons (maximally specific). Therefore any consistence criterion between items of evidence make any singleton just consistent with itself and inconsistent with all the rest so that all measures of conflict trivially particularize to the Shannon measure.

3.3.7 Maximums and minimums.

The following table summarizes some results regarding maximums and minimums.

Measure **Maximum** **Minimum**	**Type**
$U(m) = \sum_{A \subseteq X} m(A) \cdot \frac{1}{\|A\|}$ maximum $(= 1) \Leftrightarrow m$ is a probability distribution [171] minimum $(= \frac{1}{\|X\|}) \Leftrightarrow m(X) = 1$ [171]	non specificity
$U(m) = \sum_{A \subseteq X} m(A) \cdot \log\|A\|$ maximum $(= \log\|X\|) \Leftrightarrow m(X) = 1$ [97] minimum $(= 1) \Leftrightarrow m$ is a probability distribution [97]	non specificity
$U(m) = -\sum_{A \subseteq X} m(A) \cdot \log Bel(A)$ maximum $\Leftrightarrow m$ contain as many focal elements as possible such that none of them is included in any other and weights are uniformly distributed among them [32]. minimum $(= 0) \Leftrightarrow m$ has only one focal element [32].	conflict
	(to be continued)

	(continuation)
$U(m) = -\sum_{A \subseteq X} m(A) \cdot \log Pl(A)$ maximum $(= \log\|X\|) \Leftrightarrow m$ is the uniform probability distribution. minimum $(= 0) \Leftrightarrow \bigcap_{A \in X : m(A) > 0} A \neq \emptyset$ [171]	conflict
$U(m) = \sum_{A \subseteq X} m(A) \cdot \sum_{B \subseteq X} \log(m(B) \cdot \frac{\|A \cap B\|}{\|A\|})$ maximum $(= \log\|X\|) \Leftrightarrow \forall x \in X : m(\{x\}) = \frac{1}{\|X\|}$ [97] minimum $(= 0) \Rightarrow \exists x \in X : m(\{x\}) = 1$ [97]	conflict
$U(m) = \sum_{A \subseteq X} m(A) \cdot \log \frac{\|A\|^2}{\sum_{B \subseteq X} m(B) \cdot \|A \cap B\|}$ maximum $(= \log\|X\|) \Leftrightarrow m(X) = 1$ or $\forall x \in X : m(\{x\}) = \frac{1}{\|X\|}$ minimum $(= 0) \Leftrightarrow \exists x \in X : m(\{x\}) = 1$ [97]	combined
$U(m) = \sum_{A \subseteq X} m(A) \cdot \log \frac{\|A\|}{m(A)}$ maximum $\Leftrightarrow m(A) = \frac{\|A\|}{k}$ where $k = n \cdot 2^{n-1}$ [128]	combined

3.4 Other theories.

The field of imprecise probabilities comprises, but in any case is limited to, the Theory of Evidence. Many advances have been developed over the last several decades leading to a whole set of theories for representing imprecise probabilities of different levels of generality.

This section summarizes definitions of measures of uncertainty for some of these theories in order to illustrate this point.

Within the theory of Fuzzy Measures, Yager [181] suggested an extension of Shannon entropy based on the Shapley index [148] which he called Shapley entropy.

Let μ be a fuzzy measure on $X = \{x_1, \ldots, x_n\}$. Then for all $x_j \in X$ its Shapley index S_j is defined by

$$S_j = \sum_{k=0}^{n-1} \Big(\gamma_k \cdot \sum_{A \subseteq F_j : |A| = k} \big(\mu(A \cup \{x_j\}) - \mu(A) \big) \Big) \tag{3.37}$$

where

$$F_j = X - \{x_j\}$$

and

$$\gamma_k = \frac{(n-k-1)! \cdot k!}{n!}.$$

As Yager states, this index can be seen as the average increase in certitude obtained by adding element x_j to a set which does not contain it.

This index is used to define the Shapley entropy of a fuzzy measure μ as

$$U(\mu) = -\sum_{j=1}^{n} S_j \cdot \log S_j. \tag{3.38}$$

It can be shown that when μ is a probability measure, the Shapley entropy reduces to the Shannon entropy.

Abellan and Moral [1] generalized the measure of non specificity to closed convex sets of probability distributions as follows: let D be a closed convex set of probability distributions p on a domain X, g_D be the lower probability function defined by

$$\forall A \subseteq X : g_D(A) = \inf_{p \in D} \sum_{x \in A} p(x) \tag{3.39}$$

and m_D be the Möbius inverse of g_D

$$m_D = \sum_{B \subseteq A} (-1)^{|A-B|} g_D(B). \tag{3.40}$$

Then, the measure of non specificity associated with D is:

$$U(D) = \sum_{A \subseteq X} m_D(A) \cdot log|A|. \tag{3.41}$$

3.5 Summary of measures of uncertainty.

The measures of uncertainty presented are summarized in the table below.

Measure	**Type**	**Theory**
$U(A) = \log_2 \lvert A \rvert$	non specificity	classical set theory
$U(m) = \sum_{A \subseteq X, A \neq \emptyset} m(A) \cdot \log_2 \lvert A \rvert$	non specificity	evidence theory
$U(p) = -\sum_{x \in X} p(x) \cdot \log_2 p(x)$	conflict	probability theory
$U(m) = -\sum_{A \subseteq X} m(A) \cdot \log Pl(A)$	conflict	evidence theory
$U(m) = -\sum_{A \subseteq X} m(A) \cdot \log Bel(A)$	conflict	evidence theory
$U(m) = \sum_{A \subseteq X} m(A) \cdot \log \sum_{B \subseteq X} m(B) \cdot \frac{\lvert A \cap B \rvert}{\lvert B \rvert}$	conflict	evidence theory
$U(m) = \sum_{A \subseteq X} m(A) \cdot \log \sum_{B \subseteq X} m(B) \cdot \frac{\lvert A \cap B \rvert}{\lvert A \rvert}$	conflict	evidence theory
$AU(Bel) = \max_{P_{Bel}} \left[-\sum_{x \in X} p(x) \cdot \log p(x) \right]$	combined	evidence theory
$U(m) = -\sum_{A \subseteq X} m(A) \cdot \log m(A)$	conflict	evidence theory
$U(m) = -\sum_{A \subseteq X} \log Q(A)$	conflict	evidence theory
$U(A) = \int_0^{\alpha_{\max}} \frac{1}{\lvert A^\alpha \rvert} \cdot d\alpha$	non specificity	fuzzy set theory

(to be continued)

		(continuation)
$U(A) = \omega_1 \cdot b_1 - \sum_{j=2}^{n} \omega_j \cdot b_j$	non specificity	fuzzy set theory
$U_T(A) = T_1(a_1, N(T^*_{2_{j=2,\ldots,n}}\{T_3(a_j, \omega_j)\}))$	non specificity	fuzzy set theory
$U(A) = \int_0^1 \log_2 \lvert A^\alpha \rvert \cdot d\alpha$	non specificity	fuzzy set theory
$U(A) = 1 - \min_i(d(A, E_i))$	non specificity	fuzzy set theory
$U(A) = -\sum_{x \in X} \mu_A(x) \cdot \log_2 \mu_A(x) +$ $\big(1 - \mu_A(x)\big) \cdot \log_2\big(1 - \mu_A(x)\big)$	fuzziness	fuzzy set theory
$U(A) = \frac{2}{\lvert x \rvert^k} \cdot d(A, A^{\text{near}})$	fuzziness	fuzzy set theory
$U(A) = \frac{d(C,\overline{C}) - d(A,\bar{A})}{d(C,\overline{C})}$	fuzziness	fuzzy set theory
$U(A) = \frac{d(A,A^{\text{near}})}{d(A,A^{\text{far}})}$	fuzziness	fuzzy set theory
$U(A) = \otimes_{i=1}^{n} \alpha_i \cdot N(\mu_A(a_i))$	fuzziness	fuzzy set theory
$U(A) = k \cdot \sum_{x \in X} \mu_A(x) \cdot e^{1-\mu_A(x)} +$ $(1 - \mu_A(x)) \cdot e^{\mu_A(x)}$	fuzziness	fuzzy set theory
$U(m) = -\ln\big(\sum_{A,B \subseteq X : A \subseteq B} m(A)$	fuzziness	evidence theory
		(to be continued)

		(continuation)
$U(A,p) = -\sum_{x \in X} \mu_A(x) \cdot p(x) \cdot \log_2 p(X)$	combined	probability theory fuzzy set theory
$U(m) = \sum_{A \subseteq X} m(A) \cdot \log \frac{\lvert A \rvert^2}{\sum_{B \subseteq X} m(B) \cdot \lvert A \cap B \rvert}$	combined	evidence theory
$U(m) = -\sum_{A \subseteq X} m(A) \cdot \log \frac{\lvert A \rvert}{Pl(A)}$	combined	evidence theory
$U(m) = \sum_{A \subseteq X} m(A) \cdot \log \frac{\lvert A \rvert}{m(A)}$	combined	evidence theory
$\sum_{i=1}^{n} (r_i - r_{i+1}) \cdot \log i$	non specificity	possibility theory
$\sum_{i=1}^{n} (r_i - r_{i+1}) \cdot \frac{1}{i}$	non specificity	possibility theory
$-\sum_{i=1}^{n} (r_i - r_{i+1}) \cdot \log(1 - r_{i+1})$	conflict	possibility theory
$-\sum_{i=2}^{n} (r_i - r_{i+1}) \cdot \log \frac{i}{\sum_{j=1}^{i} r_j}$	conflict	possibility theory
$\sum_{i=1}^{n-1} (r_i - r_{i+1}) \cdot \log\left[1 - i \cdot \sum_{j=i+1}^{n} \frac{r_j}{j \cdot (j-1)}\right]$	conflict	possibility theory
$-\sum_{i=1}^{n} (r_i - r_{i+1}) \cdot \log(r_i - r_{i+1})$	conflict	possibility theory
$\sum_{i=2}^{n} (r_i - r_{i+1}) \cdot \log \frac{i^2}{\sum_{j=1}^{i} r_j}$	combined	possibility theory
		(to be continued)

		(continuation)
$\sum_{i=1}^{n}(r_i - r_{i+1}) \cdot \log \frac{i}{(r_i - r_{i+1})}$	combined	possibility theory

3.6 Observational entropy.

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 captured the intuitive idea that the more possible results for an experiment, the less it can be predicted. Anyway, his measure had the drawback of ignoring the probability of the events. This difficulty was overcome by Shannon [147] defining the entropy of a random variable as:

$$H(X) = -\sum_{x \in X} p(x) \cdot \log_2 p(x). \tag{3.42}$$

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 or whatever? Providing the set of events with a particular semantics requires a "further step", in the sense that we need to adapt Shannon measure in order to express random uncertainty in terms of this semantics.

In this section, this semantics will be established by defining an indistinguishability relation between 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.

Definition 3.6.1 *Let E be a T-indistinguishability operator on a set X. The observation degree of $x_j \in X$ is defined by:*

$$\pi(x_j) = \sum_{x \in X} p(x) \cdot E(x, x_j). \tag{3.43}$$

By the reflexivity of operator E, this expression can be rewritten as:

$$\pi(x_j) = p(x_j) + \sum_{x \in X | x \neq x_j} p(x) \cdot E(x, x_j). \tag{3.44}$$

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 occurrence of some element similar 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).

Proposition 3.6.2

$$\forall x \in X : 0 \leqslant \pi(x) \leqslant 1. \tag{3.45}$$

Proof 3.6.3 *Trivial.* □

Corollary 3.6.4

$$0 \leqslant \sum_{x \in X} \pi(x) \leqslant |X|. \tag{3.46}$$

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

Definition 3.6.5 *The quantity of information received by observing x_j is defined by:*

$$C(x_j) = -\log_2 \pi(x_j). \tag{3.47}$$

Definition 3.6.6 *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:*

$$HO(E, P) = \sum_{x \in X} p(x) \cdot C(x). \tag{3.48}$$

Note that if we would have defined

$$C(x_j) = \pi(x_j)$$

then

$$HO(E, P) = \sum_{x_i, x_j} p(x_i) \cdot p(x_j) \cdot E(x_i, x_j).$$

Assuming probabilistic independence we would have

$$HO(E,P) = \sum_{x_i,x_j} p(x_i,x_j) \cdot E(x_i,x_j)$$

which could be considered the expected value of the T-indistinguishability operator .

Let us suppose the following case: let $X = \{x_1, x_2\}$ be the domain, P be the probability distribution given by $p(x_1) = p(x_2) = 0.5$, and E be the classical equality relation. It is trivial to check that $HO(E,P) = 1$. This result suggests the following definition:

Definition 3.6.7 *The information received by observing an event between two equally probable and fully distinguishable, will define the unit of measure for the observational entropy: the observable bit.*

Proposition 3.6.8 *Let E, E' be two T-indistinguishability operators on X and P be a probability distribution on X.*

$$E \subseteq E' \Rightarrow HO(E,P) \geqslant HO(E',P) \tag{3.49}$$

where "$\subseteq$" is the usual containment relation between fuzzy relations as defined by Zadeh [188]:

$$E \subseteq E' \Leftrightarrow \forall x, y \in X : E(x,y) \leqslant E'(x,y). \tag{3.50}$$

Proof 3.6.9 *Trivial.* □

Corollary 3.6.10 *Let E be a T-indistinguishability operator on X, P be a probability distribution on X and $H(P)$ be the Shannon entropy of P. Then*

$$HO(E,P) \leqslant H(P) \tag{3.51}$$

obtaining the equality when E is the classical equality relation.

Proposition 3.6.11 *Let E be a T-indistinguishability operator on X, $x_i \in X$ and P be a probability distribution on X such that*

$$P(x) = \begin{cases} 1 & , x = x_i \\ 0 & , x \neq x_i \end{cases}$$

Then

$$HO(E,P) = 0. \tag{3.52}$$

Proof 3.6.12 *Trivial.* □

Proposition 3.6.13 *Let E be the T-indistinguishability operator such that $\forall x, y \in X : E(x, y) = 1$ and P be a probability distribution on X. Then*

$$HO(E, P) = 0. \tag{3.53}$$

Proof 3.6.14 *Trivial.* □

Proposition 3.6.15 *Let E be a (crisp) equivalence relation on X, P be a probability distribution on X and denoting by $H(P)$ the Shannon entropy, by X/E the quotient set and by $\overline{P}$ the induced probability distribution on X/E:*

$$\overline{P}([x]_E \in X/E) = \sum_{y \in [x]_E} p(y). \tag{3.54}$$

Then

$$HO(E, P) = H(\overline{P}). \tag{3.55}$$

3.6.1 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:

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

When working on finite domains, T-indistinguishability operators can be represented by a symmetric matrix M, where the (i, j) component takes the value $E(x_i, x_j)$ so that column (or row) i of matrix M 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" ($\approx_{x_i}$) on X, also called singleton or simply column [188] on the literature, as:

$$\forall x_j \in X : \approx_{x_i} (x_j) = E(x_i, x_j). \tag{3.56}$$

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

$$\forall r \in [0,1] : P_{\approx_{x_i}}(r) = \sum_{x \in X | \approx_{x_i}(x) = r} p(x). \tag{3.57}$$

It is trivial to check 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 its probability the sum of the probabilities of all the elements whose "similarity degree" with x_i is r. Finally, next equality holds:

$$\forall x \in X : \pi(x) = \varepsilon(G_{\approx_x}). \tag{3.58}$$

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 degree with x".

3.6.2 Simultaneous observation degree.

In this section we will introduce the concept of simultaneous observation degree. 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 similar. If events were fully distinguishable, this "overlapping" or "simultaneous observation" could not have been possible (assuming the absence of noise or error).

Before defining the simultaneous observational degree, we need to generalize definition (3.57).

Definition 3.6.16 *Let E be a T-indistinguishability operator on X, P be a probability distribution on X. $\forall A = \{x_1, \ldots, x_k\} \subseteq X$ we can define the fuzzy set "similarity degree with x_1 and ... and x_k" as:*

$$\forall x \in X : \approx_{\{x_1,\ldots,x_k\}}(x) = T(\approx_{x_1}(x), \ldots, \approx_{x_k}(x)). \tag{3.59}$$

Then, we define the random variable $G_{\{\approx_{x_1},\ldots,\approx_{x_k}\}}$ over the interval $[0,1]$ with the following probability distribution

$$\forall r \in [0,1] : P_{\{\approx_{x_1},\ldots,\approx_{x_k}\}}(r) = \sum_{x \in X | \approx_{\{x_1,\ldots,x_k\}}(x) = r} p(x). \tag{3.60}$$

Definition 3.6.17 *Let E be a T-indistinguishability operator on X and P be a probability distribution on X. The simultaneous observation degree of the subset $\{x_1, \ldots x_k\}$ is defined by:*

$$O_E(\{x_1, \ldots, x_k\}) = \varepsilon(G_{\{\approx_{x_1}, \ldots, \approx_{x_k}\}}). \tag{3.61}$$

Proposition 3.6.18 *Let E be a T-indistinguishability operator on $X = \{x_1, \ldots, x_n\}$. Then*

$$O_E(\{x_1, \ldots, x_n\}) = 1 \Leftrightarrow \forall x_i, x_j \in X : E(x_i, x_j) = 1. \tag{3.62}$$

Proof 3.6.19

$$\begin{aligned}
&O_E(\{x_1, \ldots, x_n\}) = 1 \\
&\Leftrightarrow \varepsilon(G_{\{\approx_{x_1}, \ldots, \approx_{x_n}\}}) = 1 \\
&\Leftrightarrow \sum_{x \in X | \approx_{\{x_1, \ldots, x_n\}}(x) = 1} p(x) = 1 \\
&\Leftrightarrow \exists \{x_1, \ldots, x_k\} \subseteq X : \sum_{x \in \{x_1, \ldots, x_k\}} p(x) = 1
\end{aligned}$$

such that

$$\begin{aligned}
&\forall x \in \{x_1, \ldots, x_k\} : T(\approx_{x_1}(x), \ldots, \approx_{x_n}(x)) = 1 \\
&\Leftrightarrow \forall x_i \in X : \approx_{x_i}(x) = 1 \\
&\Leftrightarrow \forall x_i \in X : E(x_i, x) = 1 \\
&\Leftrightarrow \forall x_i, x_j \in X : E(x_i, x_j) = 1
\end{aligned}$$

□

Proposition 3.6.20 *Let E be a T-indistinguishability operator on $X = \{x_1, \ldots, x_n\}$ and P be a probability distribution on X. Then*

$$O_E(\{x_1, \ldots, x_n\}) = \quad 0 \Leftrightarrow \forall x_i \in X \mid p(x_i) > 0 : \exists x_j \mid E(x_i, x_j) = 0$$

Proof 3.6.21

$$
\begin{aligned}
&O_E(\{x_1,\ldots,x_n\}) = 0\\
&\Leftrightarrow \varepsilon(G_{\{\approx_{x_1},\ldots,\approx_{x_n}\}}) = 0\\
&\Leftrightarrow \forall r \in [0,1] : p(G_{\{\approx_{x_1},\ldots,\approx_{x_n}\}} = r) = \sum_{x\in X:\approx_{\{x_1,\ldots,x_n\}}(x)=r} p(x) = 0\\
&\forall x \in X : \Big(T(\approx_{x_1}(x),\ldots,\approx_{x_n}(x)) = 0 \vee p(x) = 0\Big)\\
&\Leftrightarrow \forall x \in X \exists x_i \in X : \Big(E(x_i,x) = 0 \vee p(x) = 0\Big)\\
&\Leftrightarrow \forall x \in X \ \textit{such that}\ p(x) > 0
\end{aligned}
$$

it holds that

$$\exists x_i \in X : E(x_i,x) = 0.$$

□

3.6.3 Conditional observation degree.

In the last section we dealt with the scene in which there was disagreement between observers "equipped" with the same indistinguishability relation. Now we will consider the case in which observers have different indistinguishability abilities (each observer has his own T-indistinguishability operator). 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 similar to x_i with respect to E_A . This restriction in the set of possible events affects to the observability of observer B.

Definition 3.6.22 *Let E be a T-indistinguishability operator on X, $x_j \in X$ and P be a probability distribution on X. Then $\forall x \in X$ we define:*

$$
\begin{aligned}
P^E_{x_j}(x) &= \frac{p(x)\cdot E(x,x_j)}{\pi_E(x_j)}\\
&= \frac{p(x)\cdot E(x,x_j)}{\sum_{y\in X} p(y)\cdot E(y,x_j)}.
\end{aligned}
\tag{3.63}
$$

$P^E_{x_j}(x)$ quantifies the contribution of x to the observation degree of x_j in (E,P).

Following [29] the expression above could also be justified in probabilistic terms as follows: let us suppose a set X and a fuzzy set A on X with membership function μ_A.

Taking

$$p(A) = \sum_{x \in X} p(x) \cdot \mu_A(x) \tag{3.64}$$

in accordance with the definition of the probability of a fuzzy event given by Zadeh [187], then we could ask for the probability of the intersection of a crisp event and a fuzzy event:

$$p(A \cap \{x_i\}), x_i \in X. \tag{3.65}$$

Recalling that we are under a probabilistic setting we write:

$$p(A \cap \{x_i\}) = p(\{x_i\}) \cdot p(A|\{x_i\}).$$

Leaving aside formal details (we are just providing an intuitive interpretation) we could make the assumption:

$$p(A \cap \{x_i\}) = p(\{x_i\}) \cdot \mu_A(x_i).$$

Let us now define $p(\{x_i\}|A)$ as the probability of a crisp event conditioned to a fuzzy one as

$$p(\{x_i\}|A) = \frac{p(A \cap \{x_i\})}{p(A)} = \frac{p(x_i) \cdot \mu_A(x_i)}{\sum_x p(x) \cdot \mu_A(x)}. \tag{3.66}$$

Translating this expression to our framework we obtain:

$$p^E_{x_j}(x_i) = \frac{p(x_i) \cdot E(x_i, x_j)}{\sum_x p(x) \cdot E(x, x_i)} = p(\{x_i\}|\pi_E(x_j)). \tag{3.67}$$

The definition above could be interpreted as the probability of the crisp event x_i, conditioned to the fuzzy event of "having observed x_j in (E,P)".

Another possible interpretation is given in [29] by noticing that the use of Dempster rule to combine two belief functions, one of which being a probability measure and the other being a possibility measure, leads to an equivalent expression. Namely, let Bel_1 be a possibility measure with associated possibility distribution μ , and let Bel_2 define a probability measure p. Then

$$\forall x_i \in X : (Bel_1 \oplus Bel_2)(x_i) = \frac{\mu(x_i) \cdot p(x_i)}{\sum_{x \in X} \mu(x) \cdot p(x)} \tag{3.68}$$

which is equivalent to our definition of $P^E_{x_j}(x_i)$.

Proposition 3.6.23 *Let E be a T-indistinguishability operator on X and P be a probability distribution on X. Then*

$$\forall x_j \in X : \sum_{x \in X} P^E_{x_j}(x) = 1. \tag{3.69}$$

Proof 3.6.24 *Trivial.* □

Definition 3.6.25 *Let E, E' be two T-indistinguishability operators on X and P be a probability distribution on X. We define the conditioned observation degree of $x_i \in X$ having observed x_j in (E', P) as*

$$\pi^{E|E'}_{x_j}(x_i) = \sum_{x \in X} P^{E'}_{x_j}(x) \cdot E(x, x_i). \tag{3.70}$$

In (3.61) we defined the simultaneous observation degree. It seems natural to extend the former definition in order to consider the observation degree conditioned to a simultaneous observation of a subset of X.
As a previous step, we need to extend (3.63) in order to quantify the contribution of an element $x \in X$ to the simultaneous observation of a subset $\{x_1, \ldots, x_k\} \subseteq X$ in (E, P).

Definition 3.6.26 *Let E be a T-indistinguishability operator on X, P be a probability distribution on X and $\{x_1, \ldots, x_k\}$ be a subset of X. Then $\forall x \in X$:*

$$\begin{aligned} P^E_{\{x_1,\ldots,x_k\}}(x) = & \quad \frac{p(x) \cdot T(E(x_1, x), \ldots, E(x_k, x))}{O_E(\{x_1, \ldots, x_k\})} \\ = & \quad \frac{p(x) \cdot T(E(x_1, x), \ldots, E(x_k, x))}{\sum_{y \in X} p(y) \cdot T(E(x_1, y), \ldots, E(x_k, y))}. \end{aligned} \tag{3.71}$$

Proposition 3.6.27 *Let E be a T-indistinguishability operator on X and P be a probability distribution on X, then*

$$\forall A \in \wp(X) : \sum_{x \in X} P^E_A(x) = 1. \tag{3.72}$$

Proof 3.6.28 *Trivial.* □

Now, we can define:

Definition 3.6.29 *Let E, E' be two T-indistinguishability operators on X and P be a probability distribution on X. We define the conditioned observation degree of $x_i \in X$, having observed $\{x_1, \ldots, x_k\}$ simultaneously in (E', P) as*

$$\pi^{E|E'}_{\{x_1,\ldots,x_k\}}(x_i) = \sum_{x \in X} P^{E'}_{\{x_1,\ldots,x_k\}}(x) \cdot E(x, x_i). \tag{3.73}$$

Definition 3.6.30 *Let E, E' be two T-indistinguishability operators on X and P be a probability distribution on X. We define the simultaneous conditioned observation degree of $A \in \wp(X)$, having observed $B \in \wp(X)$ simultaneously in (E', P) as*

$$\pi^{E|E'}_{B}(A = \{a_1, \ldots, a_k\}) = \sum_{x \in X} P^{E'}_{B}(x) \cdot T(E(x, a_1), \ldots, E(x, a_k)).$$

3.6.4 Conditioned observational entropy.

Informally, the conditioned observational entropy measures how do affect the observations performed by an observer "using" a T-indistinguishability operator E' in the variability degree of the potential observations (observational entropy) of some other observer using another T-indistinguishability operator E. Of course, this influence can exist because both observers share the domain of observation.

Definition 3.6.31 *Let E, E' be two T-indistinguishability operators on X and P be a probability distribution on X. We define the observational entropy of the pair (E, P) conditioned to the observation of $x_j \in X$ in (E', P) as follows:*

$$HO_{x_j}(E \mid E', P) = -\sum_{x_i \in X} P^{E'}_{x_j}(x_i) \cdot \log_2 \pi^{E|E'}_{x_j}(x_i). \tag{3.74}$$

As we said, having observed x_j in (E', P) restricts the events that really may happened. $HO_{x_j}(E \mid E', P)$ measures the observational entropy under these new restrictions.

Definition 3.6.32 *Let E, E' be two T-indistinguishability operators on X and P be a probability distribution on X. We define the observational entropy of the pair (E, P) conditioned by the pair (E', P) as*

$$HO(E \mid E', P) = \sum_{x_j \in X} p(x_j) \cdot HO_{x_j}(E \mid E', P). \tag{3.75}$$

In other words, the conditioned observational entropy of pair (E, P) is the expected value of the observational entropy of (E, P) conditioned to the observation of all $x_j \in X$ in (E', P).

Proposition 3.6.33 *Let E' be the classical equality on X. Then for all probability distribution P on X, and for all T-indistinguishability operator E on X we have*

$$HO(E \mid E', P) = 0. \tag{3.76}$$

Proof 3.6.34

If

$$\forall x_i, x_j \in X : E'(x_i, x_j) = \begin{cases} 1 & x_i = x_j \\ 0 & otherwise \end{cases}$$

then

$$\begin{aligned} \pi_{x_j}^{E|E'}(x_i) &= \sum_{x \in X} P_{x_j}^{E'}(x) \cdot E(x, x_i) \\ &= \sum_{x \in X} \frac{p(x) \cdot E'(x, x_j)}{p(x_j)} \cdot E(x, x_i) \\ &= E(x_i, x_j). \end{aligned}$$

Therefore

$$\begin{aligned} HO_{x_j}(E \mid E', P) &= -\sum_{x_i \in X} P_{x_j}^{E'}(x_i) \cdot \log_2 \pi_{x_j}^{E|E'}(x_i) \\ &= -1 \cdot \log_2 1 = 0 \end{aligned}$$

and

$$HO(E \mid E', P) = \sum_{x_j \in X} p(x_j) \cdot HO_{x_j}(E \mid E', P) = 0.$$

□

When the conditioning T-indistinguishability operator is the classical equality relation, all observations performed on it restrict maximally the set of events that may have happened to only one event. Namely, if we observe $x_j \in X$ in the classical equality context, only x_j could have happened. Therefore, knowing which element has been observed in the pair (E', P) suppresses the variability in the restricted set of potential observations of (E, P). Consequently, the conditioned observational entropy equals zero.

Proposition 3.6.35 *Let E' be the T-indistinguishability operator such that $\forall x_i, x_j \in X : E'(x_i, x_j) = 1$, then for all T-indistinguishability operator E on X and for all probability distribution P on X, it holds:*

$$HO(E \mid E', P) = HO(E, P). \tag{3.77}$$

Proof 3.6.36

If

$$\forall x_i, x_j \in X : E'(x_i, x_j) = 1$$

then

$$\begin{aligned} \pi_{x_j}^{E|E'}(x_i) &= \sum_{x \in X} P_{x_j}^{E'}(x) \cdot E(x, x_i) \\ &= \sum_{x \in X} p(x) \cdot E(x, x_i) \\ &= \pi_E(x_i). \end{aligned}$$

Therefore

$$\begin{aligned} HO_{x_j}(E \mid E', P) &= -\sum_{x_i \in X} P_{x_j}^{E'}(x_i) \cdot \log_2 \pi_{x_j}^{E|E'}(x_i) \\ &= \sum_{x_i \in X} p(x_i) \cdot \log_2 \pi_E(x_i) \\ &= HO(E, P) \end{aligned}$$

and

$$\begin{aligned} HO(E \mid E', P) &= \sum_{x_j \in X} p(x_j) \cdot HO_{x_j}(E \mid E', P) \\ &= \sum_{x_j \in X} p(x_j) \cdot HO(E, P) \\ &= HO(E, P). \end{aligned}$$

□

Now we deal with the reverse case of the former proposition. Since the conditioning T-indistinguishability operator is maximally uncertain, all elements are fully indistinguishable. This fact causes that no restrictions on the set of events that may have happened were induced by the observations performed on (E', P). So, the only remaining restrictions are those imposed by the T-indistinguishability operator E itself. Therefore the conditioned observational entropy equals the non conditioned observational entropy.

Proposition 3.6.37 *Let E be the T-indistinguishability operator such that $\forall x_i, x_j \in X : E(x_i, x_j) = 1$. Then, for all T-indistinguishability operator E' and for all probability distribution P on X it holds:*

$$HO(E \mid E', P) = 0. \tag{3.78}$$

Proof 3.6.38

If

$$\forall x_i, x_j \in X : E(x_i, x_j) = 1$$

then

$$\pi_{x_j}^{E|E'}(x_i) = 1$$

and therefore

$$\begin{aligned} \forall x_j \in X : HO_{x_j}(E \mid E', P) = 0 \Rightarrow \\ HO(E \mid E', P) = 0. \end{aligned}$$

□

Any restriction induced by the observations performed in (E', P) doesn't "improve" our distinguishability ability, since for all restricted set of events, its members remain fully indistinguishable (due to the definition of E). So, the "improving" by conditioning averages zero.

Proposition 3.6.39 *Let E be the classical equality on X, E' be a T-indistinguishability operator on X and P be a probability distribution on X. We denote by H the Shannon entropy measure and by $P_x^{E'}$ the probability distribution defined in* (3.63)*. Then*

$$HO(E \mid E', P) = \sum_{x_j \in X} p(x_j) \cdot H(P_{x_j}^{E'}). \tag{3.79}$$

Proof 3.6.40

If

$$\forall x_i, x_j \in X : E(x_i, x_j) = \begin{cases} 1 & x_i = x_j \\ 0 & otherwise \end{cases}$$

then

$$\begin{aligned} \pi_{x_j}^{E|E'}(x_i) &= \sum_{x \in X} P_{x_j}^{E'}(x) \cdot E(x, x_i) \\ &= P_{x_j}^{E'}(x_i) \end{aligned}$$

and

$$\begin{aligned} HO_{x_j}(E \mid E', P) &= -\sum_{x_i \in X} P_{x_j}^{E'}(x_i) \cdot \log_2 \pi_{x_j}^{E|E'}(x_i) \\ &= -\sum_{x_i \in X} P_{x_j}^{E'}(x_i) \cdot \log_2 P_{x_j}^{E'}(x_i) \\ &= H(P_{x_j}^{E'}) \end{aligned}$$

and therefore

$$HO(E \mid E', P) = \sum_{x_j \in X} p(x_j) \cdot H(P_{x_j}^{E'}).$$

□

Proposition 3.6.41 *Let E, E' be two (crisp) equivalence relations on X, P be a probability distribution on X, X/E and X/E' be the quotient set of X by E and E' respectively, and $[x]_{E'}$ be the class of equivalence of x in X/E'. Then*

$$HO_{x_j}(E \mid E', P) = -\sum_{c \in X/E} \frac{p(c \cap [x_j]_{E'})}{p([x_j]_{E'})} \cdot \log_2 \frac{p(c \cap [x_j]_{E'})}{p([x_j]_{E'})} \tag{3.80}$$

where $\forall c \in \wp(X) : p(c) = \sum_{x \in c} p(x)$.

Proof 3.6.42

If both E and E' are equivalence relations, then

$$\forall x, x_j \in X : P_{x_j}^{E'}(x) = \begin{cases} \frac{p(x)}{p([x_j]_{E'})} & x \in [x_j]_{E'} \\ 0 & otherwise \end{cases}.$$

Then

$$\begin{aligned} \pi_{x_j}^{E|E'}(x_i) &= \sum_{x \in X} P_{x_j}^{E'}(x) \cdot E(x, x_i) \\ &= \sum_{x \in [x_j]_{E'}} \frac{p(x)}{p([x_j]_{E'})} \cdot E(x, x_i). \end{aligned}$$

Since when $x \notin [x_i]_E \Rightarrow E(x, x_i) = 0$ we have

$$\begin{aligned} \sum_{x \in [x_j]_{E'}} \frac{p(x)}{p([x_j]_{E'})} \cdot E(x, x_i) &= \sum_{x \in ([x_j]_{E'} \cap [x_i]_E)} \frac{p(x)}{p([x_j]_{E'})} \\ &= \frac{p([x_j]_{E'} \cap [x_i]_E)}{p([x_j]_{E'})} \\ &= p([x_i]_E \mid [x_j]_{E'}). \end{aligned}$$

Then

$$\begin{aligned} HO_{x_j}(E \mid E', P) &= -\sum_{x_i \in X} P_{x_j}^{E'}(x_i) \cdot \log_2 \pi_{x_j}^{E|E'}(x_i) \\ &= \sum_{x_i \in [x_j]_{E'}} \frac{p(x)}{p([x_j]_{E'})} \cdot \log_2 \frac{p([x_j]_{E'} \cap [x_i]_E)}{p([x_j]_{E'})} \\ &= -\sum_{c \in X/E} \frac{p(c \cap [x_j]_{E'})}{p([x_j]_{E'})} \cdot \log_2 \frac{p(c \cap [x_j]_{E'})}{p([x_j]_{E'})}. \end{aligned}$$

□

When E and E' are equivalence relations, the observational entropy of (E, P) conditioned to the observation of $x_j \in X$ in (E', P) measures how the elements of the class of equivalence of x_j in X/E' "are distributed" between the classes of X/E.

Corollary 3.6.43 *Let E, E' be two (crisp) equivalence relations on X, P be a probability distribution on X, $[x_j]_{E'}$ be the class of equivalence of x_j in X/E', then $\forall x_j \in X$ it holds:*

$$\exists c \in X/E : [x_j]_{E'} \subseteq c \Rightarrow HO_{x_j}(E \mid E', P) = 0. \tag{3.81}$$

Corollary 3.6.44 *Let E, E' be two equivalence relations on X, P be a probability distribution on X and X/E' be the quotient set of X by E', then*

$$HO(E \mid E', P) = \sum_{c' \in X/E'} p(c') \cdot HO_{x_j \in c'}(E \mid E', P) \tag{3.82}$$

where $\forall c' \in X/E' : p(c') = \sum_{x \in c'} p(x)$.

When E and E' are equivalence relations, $HO(E \mid E', P)$ is the expected value of the "distribution degree" (between the classes of X/E) for all classes in X/E'. This expression equals the heuristic function used in the construction of decision trees [133].

Corollary 3.6.45 *Let E, E' be two equivalence relations on X and P be a probability distribution on X. Then*

$$E' \subseteq E \Rightarrow HO(E \mid E', P) = 0. \tag{3.83}$$

3.6.5 Joint observational degree.

In this section we will define the joint observational degree.

Definition 3.6.46 *Let E, E' be two T-indistinguishability operators on X and P be a probability distribution on X. We define the joint observational degree of the pair (x_i, x_j) in $(E \times E', P)$ as*

$$\pi_{E\times E'}(x_i, x_j) = \pi_{E'}(x_j) \cdot \pi_{x_j}^{E|E'}(x_i). \tag{3.84}$$

This definition has a clear interpretation. Expanding the former expression we obtain:

$$\pi_{E\times E'}(x_i, x_j) = \sum_{x\in X} p(x) \cdot E'(x, x_j) \cdot E(x, x_i). \tag{3.85}$$

For the product t-norm, this expression is equivalent to:

$$\sum_{x\in X} p(x) \cdot T(E'(x, x_j), E(x, x_i)) \tag{3.86}$$

which in turn can be interpreted as the expected value of the random variable $G_{[\approx_{x_i} \wedge \approx'_{x_j}]}$, or more informally, the expected value of the similarity degree with x_i in (E, P) and with x_j in (E', P).
This interpretation suggests the next property:

Proposition 3.6.47 *Let E be a T-indistinguishability operator on $X$and O_E be the simultaneous observation degree defined in* (3.61)*. Taking the product t-norm, $\forall x_i, x_j \in X$ it holds*

$$\pi_{E\times E}(x_i, x_j) = O_E(\{x_i, x_j\}). \tag{3.87}$$

Proof 3.6.48 *Trivial.* □

3.6.6 Joint observational entropy.

Once defined the joint observational degree, we will define the joint observational entropy.

Definition 3.6.49 *Let E, E' be two T-indistinguishability operators on X and P be a probability distribution on X. $\forall x_i, x_j \in X$ we define:*

$$P_{E\times E'}(x_i, x_j) = p(x_j) \cdot P_{x_j}^{E'}(x_i). \tag{3.88}$$

Definition 3.6.50 *Let E, E' be two T-indistinguishability operators on X and P be a probability distribution on X. We define the joint observational entropy of $(E \times E', P)$ as:*

$$HO(E \times E', P) = -\sum_{x_i, x_j \in X} P_{E \times E'}(x_i, x_j) \cdot \log_2 \pi_{E \times E'}(x_i, x_j). \tag{3.89}$$

Proposition 3.6.51 *Let P be a probability distribution on X, E be a T-indistinguishability operator on X and E' be the T-indistinguishability operator such that $\forall x_i, x_j \in X : E'(x_i, x_j) = 1$. Then*

$$HO(E \times E', P) = HO(E, P). \tag{3.90}$$

Proof 3.6.52

If

$$\forall x_i, x_j \in X : E'(x_i, x_j) = 1$$

then

$$\begin{aligned} P_{E \times E'}(x_i, x_j) &= p(x_j) \cdot P_{x_j}^{E'}(x_i) \\ &= p(x_j) \cdot \frac{p(x_i) \cdot E'(x_i, x_j)}{\sum_{x \in X} p(x) \cdot E'(x, x_j)} \\ &= p(x_j) \cdot p(x_i) \end{aligned}$$

and

$$\begin{aligned} \pi_{E \times E'}(x_i, x_j) &= \sum_{x \in X} p(x) \cdot E'(x, x_j) \cdot E(x, x_i) \\ &= \sum_{x \in X} p(x) \cdot E(x, x_i) \\ &= \pi_E(x_i). \end{aligned}$$

Therefore

$$\begin{aligned} HO(E \times E', P) &= -\sum_{x_i, x_j \in X} P_{E\times E'}(x_i, x_j) \cdot \log_2 \pi_{E\times E'}(x_i, x_j) \\ &= \sum_{x_j \in X} p(x_j) \cdot \sum_{x_i \in X} p(x_i) \cdot \log_2 \pi_E(x_i) \\ &= \sum_{x_j \in X} p(x_j) \cdot HO(E, P) \\ &= HO(E, P). \qquad \square \end{aligned}$$

Proposition 3.6.53 *Let P be a probability distribution on X, E be a T-indistinguishability operator on X, E' be the classical equality on X and $H(P)$ be the Shannon entropy measure of P. Then*

$$HO(E \times E', P) = H(P). \tag{3.91}$$

Proof 3.6.54

If

$$\forall x_i, x_j \in X : E'(x_i, x_j) = \begin{cases} 1 & x_i = x_j \\ 0 & otherwise \end{cases}$$

then

$$\begin{aligned} P_{E\times E'}(x_i, x_j) &= p(x_j) \cdot P_{x_j}^{E'}(x_i) \\ &= p(x_j) \cdot \frac{p(x_i) \cdot E'(x_i, x_j)}{\sum_{x \in X} p(x) \cdot E'(x, x_j)} \\ &= \begin{cases} p(x_j) & i = j \\ 0 & otherwise \end{cases} \end{aligned}$$

and

$$\begin{aligned} \pi_{E\times E'}(x_i, x_j) &= \sum_{x \in X} p(x) \cdot E'(x, x_j) \cdot E(x, x_i) \\ &= p(x_j) \cdot E(x_j, x_i). \end{aligned}$$

Therefore

$$\begin{aligned} HO(E \times E', P) &= -\sum_{x_i, x_j \in X} P_{E\times E'}(x_i, x_j) \cdot \log_2 \pi_{E\times E'}(x_i, x_j) \\ &= \sum_{x \in X} p(x) \cdot \log_2 p(x) \\ &= H(P). \end{aligned}$$

□

Finally, we provide a theorem equivalent to the law of total entropies, but in the context of observational entropy.

Theorem 3.6.55 *Let E, E' be two T-indistinguishability operators on X and P a probability distribution on X. It holds*

$$HO(E \times E', P) = HO(E', P) + HO(E \mid E', P). \tag{3.92}$$

Proof 3.6.56 *By* (3.89) *we have:*

$$\begin{aligned} HO(E \times E', P) &= -\sum_{x_i, x_j \in X} P_{E\times E'}(x_i, x_j) \cdot \log_2 \pi_{E\times E'}(x_i, x_j) \\ &= -\sum_{x_i, x_j \in X} p(x_j) P_{x_j}^{E'}(x_i) \cdot \log_2 \pi_{E'}(x_j) \pi_{x_j}^{E|E'}(x_i) \\ &= -\sum_{x_i, x_j \in X} p(x_j) P_{x_j}^{E'}(x_i) \cdot \log_2 \pi_{E'}(x_j) \\ &= -\sum_{x_i, x_j \in X} p(x_j) P_{x_j}^{E'}(x_i) \cdot \log_2 \pi_{x_j}^{E|E'}(x_i). \end{aligned}$$

The first term can be rewritten as:

$$-\sum_{x_j \in X} p(x_j) \log_2 \pi_{E'}(x_j) \cdot \sum_{x_i \in X} P_{x_j}^{E'}(x_i).$$

By (3.69) *, $\sum_{x_i \in X} P_{x_j}^{E'}(x_i) = 1$. So we have:*

$$HO(E \times E', P) = -\sum_{x_j \in X} p(x_j) \log_2 \pi_{E'}(x_j) - \sum_{x_i, x_j \in X} p(x_j) P_{x_j}^{E'}(x_i) \cdot \log_2 \pi_{x_j}^{E|E'}(x_i).$$

The first term equals $HO(E', P)$ and the second equals $HO(E \mid E', P)$.

Therefore:

$$HO(E \times E', P) = HO(E', P) + HO(E \mid E', P).$$

Note: (3.90) and (3.91) become obvious corollaries of this theorem.

3.6.7 An example.

Given $X = \{x_1, x_2, x_3\}$, let E be the following T-indistinguishability operator (assuming the Lukasiewicz t-norm) :

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{pmatrix} 1 & 0.6 & 0.8 \\ 0.6 & 1 & 0.6 \\ 0.8 & 0.6 & 1 \end{pmatrix} \end{array}$$

and E' be the T-indistinguishability operator defined as:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.2 \\ 0.5 & 0.2 & 1 \end{pmatrix} \end{array}$$

Let P be the next probability distribution on X:

$$\begin{aligned} p(x_1) &= 0.5 \\ p(x_2) &= 0.3 \\ p(x_3) &= 0.2. \end{aligned}$$

Observational entropy of (E, P) and (E', P):

$$\begin{aligned} HO(E,P) &= -\sum_{x \in X} p(x) \cdot \log_2 \pi_E(x) \\ &= 0.54 \end{aligned}$$

$$\begin{aligned} HO(E',P) &= -\sum_{x \in X} p(x) \cdot \log_2 \pi_{E'}(x) \\ &= 1.03 \end{aligned}$$

Conditioned observational entropies:

$$\begin{aligned} HO_{x_1}(E \mid E', P) &= -\sum_{x \in X} P_{x_1}^{E'}(x) \cdot log_2 \pi_{x_1}^{E|E'}(x) \\ &= -\sum_{x \in X} \frac{p(x)E'(x,x_1)}{\pi_{E'}(x_1)} \cdot \log_2 \sum_{y \in X} P_{x_1}^{E'}(y) E(y,x_1) \\ &= 0.084 \end{aligned}$$

$$\begin{aligned} HO_{x_2}(E \mid E', P) &= -\sum_{x \in X} P_{x_2}^{E'}(x) \cdot log_2 \pi_{x_2}^{E|E'}(x) \\ &= -\sum_{x \in X} \frac{p(x)E'(x, x_2)}{\pi_{E'}(x_2)} \cdot \log_2 \sum_{y \in X} P_{x_2}^{E'}(y) E(y, x_2) \\ &= 0.137 \end{aligned}$$

$$\begin{aligned} HO_{x_3}(E \mid E', P) &= -\sum_{x \in X} P_{x_3}^{E'}(x) \cdot log_2 \pi_{x_3}^{E|E'}(x) \\ &= -\sum_{x \in X} \frac{p(x)E'(x, x_3)}{\pi_{E'}(x_3)} \cdot \log_2 \sum_{y \in X} P_{x_3}^{E'}(y) E(y, x_3) \\ &= 0.257 \end{aligned}$$

$$\begin{aligned} HO(E \mid E', P) &= \sum_{x \in X} p(x) \cdot HO_x(E \mid E', P) \\ &= 0.14 \end{aligned}$$

Joint observational entropy of $(E \times E', P)$:

$$\begin{aligned} HO(E \times E', P) &= -\sum_{x_i, x_j \in X} P_{E \times E'}(x_i, x_j) \cdot \log_2 \pi_{E \times E'}(x_i, x_j) \\ &= -\sum_{x_i, x_j \in X} p(x_j) P_{x_j}^{E'}(x_i) \cdot \log_2 \pi_{E'}(x_j) \pi_{x_j}^{E|E'}(x_i) \\ &= 1.17 \end{aligned}$$

Finally, it holds:

$$\begin{aligned} HO(E \times E', P) &= HO(E', P) + HO(E \mid E', P) \\ 1.17 &= 1.03 + 0.14 \end{aligned}$$

Chapter 4

Application to Modelling with Words.

A little inaccuracy sometimes saves tons of explanation.
Hector Hugh Munro -Saki-, "The Square Egg".

4.1 Introduction.

The main contributions of this chapter are:

- A general framework for the induction of decision trees in the presence of uncertainty.
- A new type of decision trees (observational decision trees) based on the concept of observational entropy.
- A new algorithm (FSQ) belonging to the family of sequential covering algorithms, intended to induce linguistic rules from data. A formal comparison with existing methods is also provided.
- New techniques to define the T-indistinguishability operator on a set of fuzzy subsets over X (prototypes) compatible with a given T-indistinguishability operator over the elements of X.

Recently, technological improvements and our growing ability in collecting information have raised the emergence of a great number of data sets about a very different kind of topics. But all this information will become

useless unless some efficient mechanisms are developed in order to manage this huge amount of data.

Moreover, available data sometimes happen to be pervaded with uncertainty either due to their inherent qualitative origin, or to limitations in measurement instruments. It also should be noted that lack of uncertainty may not be always a desirable property since its management could help us in obtaining simpler and more understandable representations by improving abstraction and generalization abilities. Therefore it seems reasonable asking the available methods and techniques for an adequate treatment of uncertainty.

Inductive learning has a very long tradition as a matter of concern of areas such as computer science, psychology, biology ....Nevertheless, the study and development of artificial inductive systems have been mainly focused on managing, at most, probabilistic uncertainty, neglecting other forms of uncertainty which indeed are present in the real world. With the advent of a growing interest in uncertainty beyond the probabilistic setting, a lot of work has been carried out addressing the development of methods suited to capture and manage other types of uncertainty.

As suggested in [114, 118, 139], inductive learning methods are often described in terms of the paradigm of state space search by defining the set of operators governing transitions between states, the searching procedure and the description language. The choice of the representation language should be determined in part by the task this particular method is expected to solve. So, accuracy is a key property for prediction systems while comprehensibility is agreed to be a must for knowledge extraction systems, for instance.

This chapter is intended to explore inductive methods accounting for a proper management of uncertainty and whose generated descriptions are expected to be easily interpretable.

Our proposal will follow the paradigm of "Computing with Words" which is based on the use of linguistic rules as the description language. The expression "Computing with Words" (CW) was coined by Zadeh [193] in contrast to the traditional numerical-based computation methods. As its name suggests, CW is a methodology in which words are used instead of numbers in order to mimic the kind of qualitative reasoning performed by humans. Fuzzy Set theory and Fuzzy Logic play a pivotal role since denotations for words and linguistic quantifiers are solidly rooted upon this theory through the concept of linguistic variable.

Zadeh [189] defined a linguistic variable as a fivefold structure

$$(V, T(V), U, G, M) \tag{4.1}$$

where:

- V = name of the variable

- $T(V)$ = set of atomic terms
- U = universe of discourse
- G = syntactic rules for generating valid terms
- M = semantic rules for associating the proper "meaning" (fuzzy set on U) to each valid term.

For instance, we could define the linguistic variable "Age" on the attribute "Years" (with domain $[0, 100]$) as follows:

- V = Age
- $T(V)$ = young, middle-aged, old
- $U = [0, 100]$
- G = young | middle-aged | old
- M = fuzzy sets representing the meaning of "young", "middle-aged" and "old"

Assuming this paradigm, we will present a novel method of induction of decision trees based on the measure of observational entropy which was introduced in chapter 3.

It will also be provided a method belonging to the sequential covering family of algorithms, adapted to properly managing uncertainty derived form vagueness.

Finally, some experimental results will be provided.

4.2 Previous work.

Regarding the extraction of linguistic rules from data several approaches have been proposed in the literature. As illustrative (and by no means exhaustive) examples the following may be worth mentioned.

4.2.1 Naive approximations.

A straightforward approach for the generation of linguistic rules is the method described in [75]. Having defined a linguistic variable for each attribute, every rule belonging to their cross product is generated. An obvious drawback of this approach is the computational cost involved since the number of generated rules grows exponentially with the number of attributes.

In [162] the number of rules does not depend directly on the number of attributes but on the training examples. Having also defined linguistic variables over the set of available attributes , the authors propose to generate

one rule by example, namely that having the maximum degree of truth for this particular example.

So, for a given instance $(x_1^i, \cdots, x_n^i, c^i)$ the following rule is generated "If x_1^i is A_1^i and ... and x_n^i is A_n^i then c^i is A_c^i" , where A_j^i are the linguistic labels having the maximum membership for the values x_j^i of instance i.

This procedure may lead to an inconsistent set of rules since letting each example generate one rule, is probable the existence of conflicting rules (i.e, rules having the same antecedent but different consequent). In this case the rule having maximum degree within a conflict group is selected.

The output is the result of a defuzzification process over the set of firing rules.

4.2.2 Linguistic summaries.

The concept of linguistic summary was introduced by Yager [183] in order to obtain linguistic summarizations accounting for relevant features in the set of instances.

Linguistic summaries are derived as linguistically quantified propositions with the theoretical support of works like [192].

The following patterns determine the syntax for linguistic summaries:

- "Q instances are S"
- "Q instances are $S_1 \wedge \cdots \wedge S_k$"
- "QR instances are S"
- "QR instances are $S_1 \wedge \cdots \wedge S_k$"

where S_i and R are summarizers and Q is a linguistic quantifier. Summarizers are values taken by linguistic variables (i.e, "tall" , "young" ...) while the meaning of linguistic quantifiers ("some" , "most" ...) is defined as fuzzy sets on $[0, 1]$ [192].

Linguistic summaries are validated by computing the degree of truth (ρ) of its associated pattern:

- ρ ("Q instances are S") $= \mu_Q(\frac{1}{|X|}\sum_{x \in X} \mu_S(x))$
- ρ ("Q instances are $S_1 \wedge \ldots \wedge S_k$") $= \mu_Q(\frac{1}{|X|}\sum_{x \in X} T(\mu_{S_1}(x), \ldots, \mu_{S_k}(x)))$
- ρ ("Q R instances are S") $= \mu_Q(\frac{\sum_{x \in X} T(\mu_R(x), \mu_S(x))}{\sum_{x \in X} \mu_R(x)})$
- ρ "Q R instances are $S_1 \wedge \ldots \wedge S_k$" $= \mu_Q(\frac{\sum_{x \in X} T(\mu_R(x), \mu_{S_1}(x), \ldots, \mu_{S_k}(x))}{\sum_{x \in X} \mu_R(x)})$

In addition, Yager also introduces the notion of informativeness for linguistic summaries. Informativeness is estimated measuring the difficulty of "reconstructing" the original data set from which the summary was derived. Consequently, the informativeness of a given summary S is defined as the specificity of the set of data sets consistent with S ($CON(S)$) as:

$$Informativeness(S) = \frac{1}{|CON(S)|} \tag{4.2}$$

In [102] a partial order is defined for each attribute in the set of its corresponding linguistic values. This order is intended to model the prior knowledge about the structure of the domains.

Descriptions are represented as propositions in conjunctive form. A breadth search is performed (from general to specific descriptions) pruning the descriptions whose support is below a predefined threshold. The correctness of this pruning strategy derives from the fact that if the support of a given description D does not reach a given threshold, no specialization of D will do.

Informativeness is defined based on Hartley measure, so that the informativeness of a given description S quantifies the reduction of uncertainty when characterizing a particular data set (the one from which D was generated) resulting from knowing that the set of data sets have to be consistent with S.

$$Informativeness(S) = \log \frac{|TOT|}{|CON(S)|} \tag{4.3}$$

where TOT is the set of all possible data sets and $CON(S) \subseteq TOT$ is the set of data sets consistent with S.

4.2.3 Sequential covering.

The family of methods known as sequential covering algorithms [19, 113] are intended to produce a disjunctive set of rules which "cover" a given subset of instances (target concept) of the training set by following the next scheme: first, a rule is generated covering a "portion" of the target concept. We require this rule having high accuracy, but not necessarily high coverage. By high accuracy we mean the predictions it makes should be correct. By accepting low coverage, we mean it needs not make predictions for every training example [113]. Then, instances covered by this rule are removed from the training set. This procedure can be iterated as many times as desired to learn a disjunctive set of rules that together cover any desired fraction of the training set.

4.2.4 Decision trees.

Decision trees, since their formal appearance within the context of inductive learning [133] have become one of the most relevant paradigm of machine learning methods. The main reason for this wide-spreading success lies in their proved applicability to a broad range of problems, in addition to appealing features as the readability of the knowledge represented in the tree. Therefore, a lot of work have been carried out from Quinlan's TDID3 algorithm in order to extend the applicability to domains beyond the categorical ones and achieve further improvements. In this line, many approaches dealing with continuous-valued attributes have been proposed ([16, 135, 107]). Also, alternative measures to the classical Shannon entropy measure [147] for attribute selection have been devised, like Gini's test [16], Kolmogorov-Smirnoff distance [157], distance between partitions [108], contrast measures [159], ...

Another important point has been providing decision tree induction algorithms with a more flexible methodology in order to cope with other sources of uncertainty beyond the probabilistic type. Indeed, when we face real problems we should overcome the limitations of the probabilistic framework by furnishing existing methods, so that other well-known types of uncertainty could be managed.

Some generalizations have been proposed to achieve this goal.

Probabilistic decision trees.

Coping with missing or uncertain attribute values where uncertainty is represented by probability distributions led Quinlan to develop probabilistic decision trees [134].

Fuzzy decision trees.

Generalization and interpolation properties of fuzzy sets together with their less sensitivity to small changes in input attribute values make them specially suitable for classification tasks. Fuzzy decision trees ([82], [163], [185], [156], [2]) benefit from those aspects to provide a flexible framework for inducing linguistic rules.

In [185] the selection of the splitting attribute is based on computing the specificity of the fuzzy partition induced by the values of the corresponding attribute.

For a given instance u, let $E(u)$ be the accumulated evidence in node k computed as the conjunction of membership degress for all linguistic labels from k up to the root.

Then, for a given remaining (non used) attribute A, for all linguistic value A_i of A and for all class C_i, the degree of possibility of class C_i given the evidence $E \wedge A_i$ is defined as

$$\pi(C_i|E \wedge A_i) = \frac{\sum_{u \in U} \mu_{C_i}(u) \wedge \mu_{A_i}(u) \wedge E(u)}{\sum_{k=1}^{j} \sum_{u \in U} \mu_{A_i}(u) \wedge C_k(u) \wedge E(u)} \tag{4.4}$$

where U is the set of instances and j the number of classes. The main difference with respect to other methods is taking the value π as a possibility degree instead of the usual probabilistic interpretation.

Therefore for value A_i, the following possibility distribution is obtained:

$$\pi_{A_i} = (\frac{\pi(C_1|E \wedge A_i)}{max_{1 \leqslant k \leqslant j}\{\pi(C_k|E \wedge A_i)\}}, \ldots, \frac{\pi(C_j|E \wedge A_i)}{max_{1 \leqslant k \leq j}\{\pi(C_k|E \wedge A_i)\}}). \tag{4.5}$$

U-uncertainty measure is then computed for the possibility distribution above:

$$g(\pi_{A_i}) = \sum_{i=1}^{j} (\pi_i - \pi_{i+1}) \cdot \log i. \tag{4.6}$$

Finally, $g(\pi_{A_i})$ values are averaged for linguistic values A_i of attribute A, yielding:

$$g(A) = \sum_{A_i \in A} \frac{\sum_{u \in U} \mu_{A_i}(u) \wedge E(u)}{\sum_{A_j \in A} \sum_{u \in U} \mu_{A_j}(u) \wedge E(u)} \cdot g(\pi_{A_i}). \tag{4.7}$$

The attribute minimizing $g(A)$ will be selected.

In [82] the traditional approach of minimizing the entropy is followed. This method adapts the classical procedure in order to properly deal with linguistic variables.

With the nomenclature introduced before, for all linguistic value A_i of attribute A and for all class C_i, the probability of C_i conditioned to evidence $E \wedge A_i$ is defined as:

$$p(C_i|E \wedge A_i) = \frac{\sum_{u \in U} \mu_{C_i}(u) \wedge \mu_{A_j}(u) \wedge E(u)}{\sum_{k=1}^{j} \sum_{u \in U} \mu_{A_i}(u) \wedge \mu_{C_k}(u) \wedge E(u)}. \tag{4.8}$$

For each value A_i, the entropy of the distribution of instances compatible with $E \wedge Ai$ with respect to their distribution among the set of classes is computed by:

$$H(C|E \wedge A_i) = -\sum_{k=1}^{j} p(C_j|E \wedge A_i) \cdot \log p(C_j|E \wedge A_i). \tag{4.9}$$

Finally, entropy for attribute A is computed averaging the entropy for each linguistic value A_i of A as

$$H(A) = \sum_{A_i \in A} \frac{\sum_{u \in U} \mu_{A_i}(u) \wedge E(u)}{\sum_{A_j \in A} \sum_{u \in U} \mu_{A_j(u)} \wedge E(u)} \cdot H(C|E \wedge A_i). \tag{4.10}$$

The selected attribute will be the one minimizing $H(A)$.

Belief decision trees.

Belief measures [145] provide a mechanism to express and deal with subjective judgments in a much more flexible way than probability, offering tools for properly handling ignorance and combining several pieces of evidence. Hence it was advisable to integrate the advantages of belief functions and decision trees, resulting in the belief decision trees approach [37, 7].

In [37] a procedure for inducing decision trees when evidences about the class attribute C are represented as basic probability assignments over the domain of C (set of possible classes, assuming that C is a symbolic attribute) is introduced.

Two approaches are presented in order to select the best splitting attribute.

Averaging approach.

The compatibility between a given class C_i and a given instance representing evidence about its class membership in the form of a *bpa* m over the set of classes is computed using the definition of the pignistic probability [150] as

$$\pi(C_i|m) = \sum_{A \subseteq C : C_i \in A} \frac{1}{|A|} \cdot \frac{m(A)}{1 - m(\emptyset)}. \tag{4.11}$$

Then, probability of class C_i in node n is defined as the average of values $\pi(C_i|m)$ over the set of instances (I_n) "arriving" to node n:

$$p(C_i) = \frac{1}{|I_n|} \cdot \sum_{i \in I_n} \pi(C_i|m). \tag{4.12}$$

Once defined these probabilities, the attribute minimizing the entropy is selected.

Conjunctive approach.

In this case the selected attribute is that partitioning the set of instances in such a way that evidence (*bpa*) about class membership in each instance is similar to those of instances belonging to the same partition group.

For this purpose an intra-group measure of distance is defined. The attribute minimizing the averaged intra-group distance over its set of values, is selected.

Stopping criteria.

Selection and splitting processes are iterated until one of the following stopping criteria is fulfilled for any leaf node n:

- Only one instance reaches n.
- Instances reaching n represent all the same evidence (*bpa*).
- The set of remaining splitting attributes is empty.
- There is no gain in partition the set of instances "arriving" to n.

4.2.5 Others.

Other group is composed by methods which make use of some biological inspired mechanism (neural networks [100], genetic algorithms [68], ant colonies [18], ...) in order to induce a set of linguistic rules.

Finally, there are some methods which do not fit in any of the previous groups like the induction of linguistic functional dependencies which, based on the concept of fuzzy graph [193], have the objective of grasping existing functional dependencies and expressing them in a qualitative manner. References [162, 36] are examples of this approximation.

4.3 A general framework for induction of decision trees under uncertainty.

A decision tree can be viewed as a representation of a procedure to determine the classification of an object. Any specific decision tree based technique should deal basically with two main concerns, namely, how to build the tree out of a set of examples, and how it is going to be used; corresponding to the definition of a building procedure and an inference algorithm, respectively.

The building procedure usually follows the basic scheme by Quinlan [133] based on a top down strategy (top down induction of decision tree (TDIDT)) which proceeds by successively partitioning the training set as detailed in the procedure below:

1. Place the initial data on the root.
2. Select the best attribute from the set of non used attributes and mark it as used.
3. Create new child nodes according to the partition induced by the selected attribute.
4. For each newly generated child node iterate step 2 unless any stopping criterion holds. In this case mark current node as a leaf and compute its associated label.

On the other hand, the inference process aims at classifying a new instance by traversing down the proper branch of the tree until its corresponding leaf has been reached. In order to cope with uncertainty within this process some steps must be adapted. Indeed, since the partitioning strategy does not already define an equivalence relation, an instance can follow several paths down in the tree to a certain degree and, consequently, several leaves could be reached whose labels should be combined to produce a classification. Hence, the inference algorithm should involve the next two steps:

- Computing the set of leaves reached by the instance to classify.
- Combining their associated labels to produce the output classification.

Our claim is that any decision tree based method admits a decomposition in terms of the points we are going to describe in the following subsections, so that a given method should be describable by means of a concrete configuration defined over them. Let us examine these points in more detail.

4.3.1 Structure of the training set.

One major requirement when defining a general framework for induction of decision trees should be to integrate and manage different types and representations of uncertainty in an homogeneous way. So, the framework should allow us to deal with attributes pervaded with different kinds of uncertainty described in terms of the following training set structure.

Let A be the set of attributes, let $c \in A$ (class attribute) be a distinguished attribute providing information about the class to which each instance belongs. Let E be the set of instances where, for all instance $e \in E$ and for all attribute $a \in A$, e_a is the available evidence (possibly uncertain) belonging to instance e about the value $v \in domain(a)$ taken by attribute a.

On the other hand, for each attribute $a \in A$ a set of linguistic labels $L_a = \{a_1, \ldots, a_{|L_a|}\}$ whose meaning are fuzzy sets in the corresponding domain ($domain(a)$) is defined.

These labels will "decorate" nodes and edges of the tree and shall make up the representation language for expressing the linguistic classification rules derived from the tree.

Now the question turns into how to manage different representations of evidences in a consistent way. A solution to this problem could be performing on the initial training set a transformation similar to the so called "binning" [14] in the classical setting. This transformation expands each attribute column of the initial training set in so many columns as the number of linguistic labels ($|L_a|$) defined for the attribute.

The cell of the expanded training set corresponding to linguistic label $a_i \in L_a$ and instance e will contain the compatibility degree between evidence e_a and linguistic label a_i given by a proper compatibility measure in such a way that these degrees could be interpreted as the approximation of evidence e_a in the linguistic label space of attribute a (L_a).

Those values are expected to quantify the degree of compatibility between evidences and labels, even when both of them may be expressed in different formal theories.

Despite the particular theory used to model a given piece of evidence or label, sharing a common domain of discourse allows to define the degree of compatibility as the possibility degree of a given label conditioned to the occurrence of a particular piece of evidence (or the expected value, for situations when probabilistic uncertainty is involved in the representation of evidence.)

Table 4.1 summarizes definitions of the compatibility degree for different types of evidences and labels. Since the use of linguistic values is tacitly assumed for the set of labels, their meaning is defined by fuzzy sets accordingly. On the other side, evidences may use different representations, ranging from singleton crisp sets to fuzzy basic probability assignments.

Let $X = \{x_1, \ldots, x_n\}$ be the domain of discourse.

Table 4.1: Compatibility measures.

	Crisp Singleton Set ($l \in X$)	**Crisp Interval** ($L \subseteq X$)	**Fuzzy Set** ($\mu_L \in [0,1]^X$)
Crisp Singleton Set ($e \in X$)	$\begin{cases} 1 & e = l \\ 0 & \text{otherwise.} \end{cases}$	$\begin{cases} 1 & e \in L \\ 0 & \text{otherwise.} \end{cases}$	$\mu_L(e)$
Crisp Set ($E \subseteq X$)	$\begin{cases} 1 & l \in E \\ 0 & \text{otherwise.} \end{cases}$	$\begin{cases} 1 & E \cap L \neq \emptyset \\ 0 & \text{otherwise.} \end{cases}$	$\sup_{e \in E} \mu_L(e)$
Fuzzy Set ($\mu_E \in [0,1]^X$)	$\mu_E(l)$	$\sup_{l \in L} \mu_E(l)$	$\sup_{x \in X} T(\mu_L(x), \mu_E(x))$
Probability Distribution (p_E **on** X)	$p_E(l)$	$\sum_{l \in L} p_E(l)$	$\sum_{x \in X} p_E(x) \cdot \mu_L(x)$
bpa (m_E **on** X)	$\sum_{A \subseteq X : l \in A} m_E(A)$	$\sum_{A \subseteq X : A \cap L \neq \emptyset} m_E(A)$	$\sum_{A \subseteq X} m_E(A) \cdot \sup_{x \in A} \mu_L(x)$
Fuzzy bpa (m_E^* **on** X)	$\sum_{\phi \in [0,1]^X} m_E^*(\phi) \cdot \phi(l)$	$\sum_{\phi \in [0,1]^X} m_E^*(\phi) \cdot \sup_{x \in L} \phi(x)$	$\sum_{\phi \in [0,1]^X} m_E^*(\phi) \cdot sup_{x \in X} T(\phi(x), \mu_L(x))$

A few remarks are worth to be made. Let (i, j) be the coordinates of cell in row i and column j in the table above. As stated previously, letting $L = \{L_1, \ldots, L_p\}$ be the set of labels and E be a given evidence, the following fuzzy set:

$$\psi : L_i \longrightarrow \text{compatibility}(L_i, E) \tag{4.13}$$

could be taken as the approximation of evidence E in the linguistic label space L.

With this interpretation, the compatibility degree defined in cell $(2, 2)$ equals the upper approximation of crisp set E in L as defined in [129]. Definitions in cells $(3, 2)$ and $(3, 3)$ correspond to rough fuzzy set and fuzzy rough set upper approximations of evidence respectively, as described in [34], and cell $(4, 3)$ is equivalent to the definition of probability of a fuzzy set introduced by Zadeh [187].

For the case at hand, evidences play the same role as generalized constraints do in the paradigm of Computing with Words [193]. Further extensions will not be addressed in this dissertation as, for instance, representation and management of second type probabilistic information [191] intended to model the situation in which available evidence only informs about the probability of occurrence of a subset of the frame of discernment, being this evidence compatible with many representations of "first order" type (probability distributions, ...).

Another possibly extension could be allowing the compatibility degree be fuzzy valued instead of scalar, as assumed in this section.

4.3.2 Node membership function.

As pointed out previously, TDIDT techniques rely on a "divide-and-conquer" paradigm by continuously partitioning the remaining set of instances as an effect of adding new constraints to be fulfilled. These new constraints come from the set of linguistic labels of the attribute selected to partition the set of instances.

In the classical setting, each partition defines an equivalence relation and the degree of membership of a given instance to each class of the quotient set is defined by the boolean conjunction of the set of constraints appearing when following the proper path up to the root.

When facing with uncertain evidences, "testing" by an attribute label does not usually produce a boolean answer. Instead of this, a compatibility degree between the evidence and the label should be managed.

Let N be the set of nodes, $n \in N$ be a given node and $R = \{r_1, \ldots, r_p\}$ be the set of constraints belonging to the path going from the root node to n. The fuzzy set $\mu_n : E \longrightarrow [0, 1]$ is defined over the set of instances as

$$\forall e \in E : \mu_n(e) = g(r_1(e), \ldots, r_p(e)) \tag{4.14}$$

where g is a conjunctive aggregation operator (usually a t-norm) and $r_i(e)$ the compatibility degree between instance e and the linguistic label corresponding to restriction r_i.

4.3.3 Attribute selection.

All along this section we will assume the following nomenclature:

- E: set of instances and $e \in E$ a particular instance.
- N: set of nodes and $n \in N$ a particular node.
- A: set of attributes and $a \in A$ a particular attribute.
- L_a: set of linguistic labels defined for attribute a and $a_i \in L_a$ a particular linguistic label of attribute a.
- Let $a_i \in L_a$ be a linguistic label of attribute a and $n \in N$, we note by $(n|a_i)$ the node whose associated set of constraints[1] is the result of appending the constraints leading to node n with the constraint $a = a_i$.
- Let $n \in N$, we note by μ_n the node n membership function as defined in (4.14).
- Let $c_i \in L_c$ be a linguistic label defined for the class attribute c, we define $\nu_{c_i} : E \longrightarrow [0, 1]$ as

  $$\forall e \in E : \nu_{c_i}(e) = m_c(e_c, c_i)$$

  where m_c is the compatibility measure between evidences and labels for class attribute c.
- T: set of normalization functions (usually probabilistic or possibilistic normalization) and $t \in T$ a particular normalization function.
- F: set of uncertainty measures and $f \in F$ a particular uncertainty measure function.
- G: set of aggregation operators[2] and $g \in G$ a particular aggregation operator. We will also use the notation g^{SET}_{NODE} where SET is an index set referencing the values to be aggregated, and $NODE$ is the referential set of instances.

[1] Defined as the constraints belonging to the path going from node n up to the root.
[2] For an in-depth study of aggregation operators the reader is referred to [109, 17]

In the classical picture we have several alternatives for quantifying the uncertainty associated to a given node. When shifting to the non classical picture, the repertory of measures goes even wider, reflecting the broadly accepted fact that there exist different kinds of uncertainty beyond the probabilistic one and consequently, some well-established measures have been developed to cope with it.

Nevertheless, a concrete realization of the general framework is not only particularized by the uncertainty measure but also by the selected aggregation operators and by the order in which they are applied. We have basically two possibilities depending on whether we first apply the uncertainty measure (f) to each instance before the resulting values are aggregated at the corresponding node ("horizontal folding"), or we perform it the opposite way ("vertical folding"). Therefore, when considering attribute a as candidate to become selected as the current branching attribute we have basically the two schemes below as options for computing uncertainty in node n.

1. Horizontal folding:

$$UNC(n) = g_n^{a_i \in A_L}(g_{(n|a_i)}^{e \in E}(f(t(\nu_{c_1}(e), \ldots, \nu_{c_{|L_c|}}(e)))))$$

2. Vertical folding:

$$UNC(n) = g_n^{a_i \in A_L}(f(t(g_{(n|a_i)}^{e \in E}(\mu_{(n|a_i|c_1)}(e)), \ldots, g_{(n|a_i)}^{e \in E}(\mu_{(n|a_i|c_{|L_c|})}(e)))))$$

4.3.4 Inference algorithm.

Coping with uncertainty makes that in contrast to the classical case, neither a unique leaf is usually reached nor a single class label could be tagged on a given leaf, making it necessary to define a procedure for combining the classifications associated to the set of reached leaves.

Therefore, and arguing in a similar way that of the previous subsection, two basic schemes for combining these classifications are shown below, where u is the instance to classify, h a particular leaf and s a "collapsing" function (i.e, a function intended to provide a precise single value from an imprecise classification (majority policies, defuzzification methods, ...)).

1.

$$CLASIF(u) = s(g_u^{h \in H}(g_h^{e \in E}(\mu_{(h|c_1)}(e))), \ldots, g_u^{h \in H}(g_h^{e \in E}(\mu_{(h|c_{|L_c|})}(e))))$$

2.

$$CLASIF(u) = g_u^{h \in H}(s(g_h^{e \in E}(\mu_{(h|c_1)}(e)), \ldots, g_h^{e \in E}(\mu_{(h|c_{L_c})}(e))))$$

4.3.5 Characterization of existing families of methods.

Tables 4.3 and 4.5 contain the characterization of representative methods belonging to families of procedures for inducing decision trees in the presence of uncertainty. These families can be grouped in three main categories:

1. Fuzzy probabilistic decision trees (represented by [82]): adaptation to the fuzzy case of classical entropy-based induction algorithms.

2. Possibilistic decision trees (represented by [185, 140]: use of possibility theory is extensively made in order to build the decision tree.

3. Belief decision trees (represented by [37]): exploit representational power and evidence combination rules provided by belief functions theory in order to model a subjective belief approach to the problem of growing decision trees.

Table 4.3: Methods characterization

	Classical id3 [133]	**Possibilistic id3 [185]**	**Possibilistic id3 [140]**
attribute evidences	nominal	fuzzy	nominal
class attribute evidences	nominal	fuzzy	not determined
attribute labels	nominal	fuzzy	nominal
class attribute labels	nominal	fuzzy	nominal
attribute compatibility measure (m)	crisp equality	not determined (values are given)	crisp equality
class attribute compatibility measure (m_c)	crisp equality	not determined (values are given)	not determined (values are given)
restrictions aggregation op $(g_e^{r_i \in R})$	boolean and	min t-norm	boolean and
node instance aggregation op $(g_n^{e \in E})$	sum	sum	max t-conorm
leaf instance aggregation op $(g_h^{e \in E})$	sum	sum	max t-conorm
leaves aggregation op $(g_u^{h \in H})$	(just one leaf reached)	max t-conorm	(just one leaf reached)
labels aggregation op $(g_n^{a_i \in A_L})$	weighted mean	weighted mean	max-min
normalization function (t)	probabilistic normalization	possibilistic normalization	not determined
collapsing function (s)	majority class	class with highest membership	class with less fuzziness
uncertainty measure (f)	entropy	U-uncertainty	U-uncertainty

Table 4.5: characterization table (cont)

	Fuzzy probabilistic id3 [82]	**Belief id3 [37]**	**Observational id3 [62]**
attribute evidences	numerical	nominal	nominal + indistinguishability
class attribute evidences	numerical	bpa	nominal + indistinguishability
attribute labels	fuzzy set	nominal	nominal + indistinguishability
class attribute labels	fuzzy set	nominal	nominal + indistinguishability
attribute compatibility measure (m)	$Pos(label\|evid) = \mu_{label}(evid)$	crisp equality	conditioned observation degree
class attribute compatibility measure (m_c)	$Pos(label\|evid) = \mu_{label}(evid)$	pignistic probability	t-norm
restrictions aggregation op $(g_e^{r_i \in R})$	t-norm	boolean and	t-norm
node instance aggregation op $(g_n^{e \in E})$	sum	sum	sum
leaf instance aggregation op $(g_h^{e \in E})$	sum	bpa averaging function	sum
leaves aggregation op $(g_u^{h \in H})$	weighted mean	disjunctive rule combination	(just one leaf reached)
labels aggregation op $(g_n^{a_i \in A_L})$	weighted mean	weighted mean	weighted mean
normalization function (t)	probabilistic normalization	probabilistic normalization	probabilistic normalization
collapsing function (s)	defuzzification	pignistic decision	majority class
uncertainty measure (f)	entropy	entropy	observational entropy

Table 4.7: Original data set.

instance	*outlook*	*temperature*	*windy*	*play*
1	sunny	hot	false	volley
2	sunny	hot	true	swimming
3	overcast	hot	false	tennis
4	rainy	mild	false	football
5	rainy	cool	true	football
6	overcast	cool	true	football
7	sunny	mild	false	tennis
8	sunny	mild	true	swimming
9	overcast	hot	false	tennis
10	rainy	mild	true	football

4.4 Observational decision trees.

In this section we will introduce a new approach to building a decision tree addressing the case when uncertainty arises as a consequence of having defined indistinguishability relations on the domains of the attributes used to describe the set of instances. Existing methods make the assumption that different events are perfectly distinguishable from each other when measuring, for instance, node's impurity (for entropy-based methods). In front of this restrictive assumption we advocate for a more realistic setting in which decision maker's discernment abilities should be taken into account, and therefore, impurity should be measured accordingly to his frame of discernment.

4.4.1 Induction algorithm.

We have already introduced the concept of observational entropy in chapter 3. Let us see how to use it for the task of building a decision tree from a set of examples. The problem could be posed as follows: Let $\{A_1, \ldots, A_n, C\}$ be a set of nominal attributes (where the classes of C form the classification we want to learn), with domains $D_i = \{v_{i_1}, \ldots v_{i_{mi}}\}$ and $D_c = \{v_{c_1}, \ldots, v_{c_{mc}}\}$. Let $E \subseteq D_1 \times \ldots \times D_n \times D_c$ be the set of instances, and for each attribute A_i we consider a T-indistinguishability operator E_{A_i} and a probability distribution P_{A_i} defined on the domain of A_i. Let us illustrate the above definitions with the example of tables 4.7 and 4.8.

$E_{Outlook} =$

	sunny	*overcast*	*rainy*
sunny	1	0	0
overcast	0	1	0.5
rainy	0	0.5	1

$E_{Temp} =$

	hot	*mild*	*cool*
hot	1	0.5	0.5
mild	0.5	1	0.5
cool	0.5	0.5	1

$E_{Play} =$

	swimming	*football*	*tennis*	*volley*
swimming	1	0	0	0
football	0	1	0.25	0.25
tennis	0	0.25	1	1
volley	0	0.25	1	1

$E_{Windy} =$

	true	*false*
true	1	0
false	0	1

Table 4.8: T-Indistinguishability operators.

$$
\begin{aligned}
D_{Outlook} &= \{sunny, overcast, rainy\} \\
D_{Temperature} &= \{hot, mild, cool\} \\
D_{Windy} &= \{true, false\} \\
D_{Play} &= \{swimming, tennis, football, volley\}
\end{aligned}
$$

In order to simplify, we assume that the probability distribution associated to each attribute of the example is defined as the uniform distribution on the corresponding domain. Generalizing this assumption is straightforward.

Next we present an algorithm for building a decision tree based on the observational entropy. The procedure could be summarized in the following points:

"Expanded" data set.

From the original data set we create its associated "expanded" data set. For all instances, we compute the compatibility between each label and the evidence (represented by the instance) by computing the conditioned observational degree between the given label and the proper component of the instance.

As an example, let us describe how to compute the compatibility between the value *overcast* for attribute *outlook* and the label *rainy*. We want to compute:

$$\pi_{overcast}(rainy) = \sum_{x \in \{sunny, overcast, rainy\}} P_{overcast}(x) \cdot E_{Outlook}(x, rainy)$$

where

$$P_{overcast}(sunny) = \frac{E_{Outlook}(sunny, overcast)}{\sum_{x \in \{sunny, overcast, rainy\}} E_{Outlook}(x, overcast)}$$

and

$$P_{overcast}(overcast) = \frac{E_{Outlook}(overcast, overcast)}{\sum_{x \in \{sunny, overcast, rainy\}} E_{Outlook}(x, overcast)}$$

and

$$P_{overcast}(rainy) = \frac{E_{Outlook}(rainy, overcast)}{\sum_{x \in \{sunny, overcast, rainy\}} E_{Outlook}(x, overcast)}$$

and therefore

$$\begin{aligned} \pi_{overcast}(rainy) &= 0 + \frac{1}{3} + \frac{1}{3} \\ &= \frac{2}{3}. \end{aligned}$$

The resulting "expanded" data set is depicted in table 4.9.

Table 4.9: Expanded data set

instance	*sunny*	*overcast*	*rainy*	*hot*	*mild*	*cool*	*true*	*false*	*swimming*	*tennis*	*football*	*volley*
1	1	0	0	0.7	0.6	0.6	0	1	0	0.9	0.9	0.1
2	1	0	0	0.7	0.6	0.6	1	0	1	0	0	0
3	0	0.8	0.6	0.7	0.6	0.6	0	1	0	0.9	0.9	0.1
4	0	0.6	0.8	0.6	0.7	0.6	0	1	0	0.5	0.5	0.7
5	0	0.6	0.8	0.6	0.6	0.7	1	0	0	0.5	0.5	0.7
6	0	0.8	0.6	0.6	0.6	0.7	1	0	0	0.5	0.5	0.7
7	1	0	0	0.6	0.7	0.6	0	1	0	0.9	0.9	0.1
8	1	0	0	0.6	0.7	0.6	1	0	1	0	0	0
9	0	0.8	0.6	0.7	0.6	0.6	0	1	0	0.9	0.9	0.1
10	0	0.6	0.8	0.6	0.7	0.6	1	0	0	0.5	0.5	0.7

Computing probabilities of observing events in a node n.

Values contained in the expanded data set will be used to compute the compatibility degree (COM) between a conjunction of restrictions and the evidence represented by a given instance e:

$$COM(A_i = v_{i_j} \wedge \ldots \wedge A_k = v_{k_l}|e) = T(\pi(v_{i_j}|e_{A_i}), \ldots, \pi(v_{k_l}|e_{A_k})).$$

Let n be a given node belonging to the current tree (the one which has been grown up to now) and let R be the conjunction of the restrictions found in the path going from the root of the tree to node n. We define the probability of observing label v_{i_j} of attribute A_i in node n as:

$$P_N(A_i = v_{i_j}) = \frac{\sum_{e \in E} COM((R \wedge A_i = v_{i_j})|e)}{\sum_{v_i \in D_i} \sum_{e \in E} COM((R \wedge A_i = v_i)|e)}.$$

Selecting branching attribute.

In the previous point we have provided a method for computing the probabilities of observing the labels for all the attributes in a given node n. These values will allow us to select the best attribute to partition data "arriving" at node n (fulfilling the restrictions leading to node n). Given a node n, we compute (for all non previously selected attributes) the observational entropy of class attribute (C) conditioned to a given remaining attribute A_i in the following manner:

$$HO(C|A_i) = \sum_{v_i \in D_i} P_n(A_i = v_i) \cdot HO(C|A_i = v_i)$$

where

$$HO(C|A_i = v_i) = -\sum_{v_c \in D_C} P_{n \wedge (A_i = v_i)}(c = v_c) \cdot \log_2 \sum_{w_c in D_C} P_{n \wedge (A_i = v_i)}(C = w_c) \cdot E_C(w_c, v_c)$$

where $P_{n \wedge (A_i = v_i)}$ are the probabilities measured in each one of child nodes of n induced by partition data arriving at node n, in accordance with the set of labels of attribute A_i.

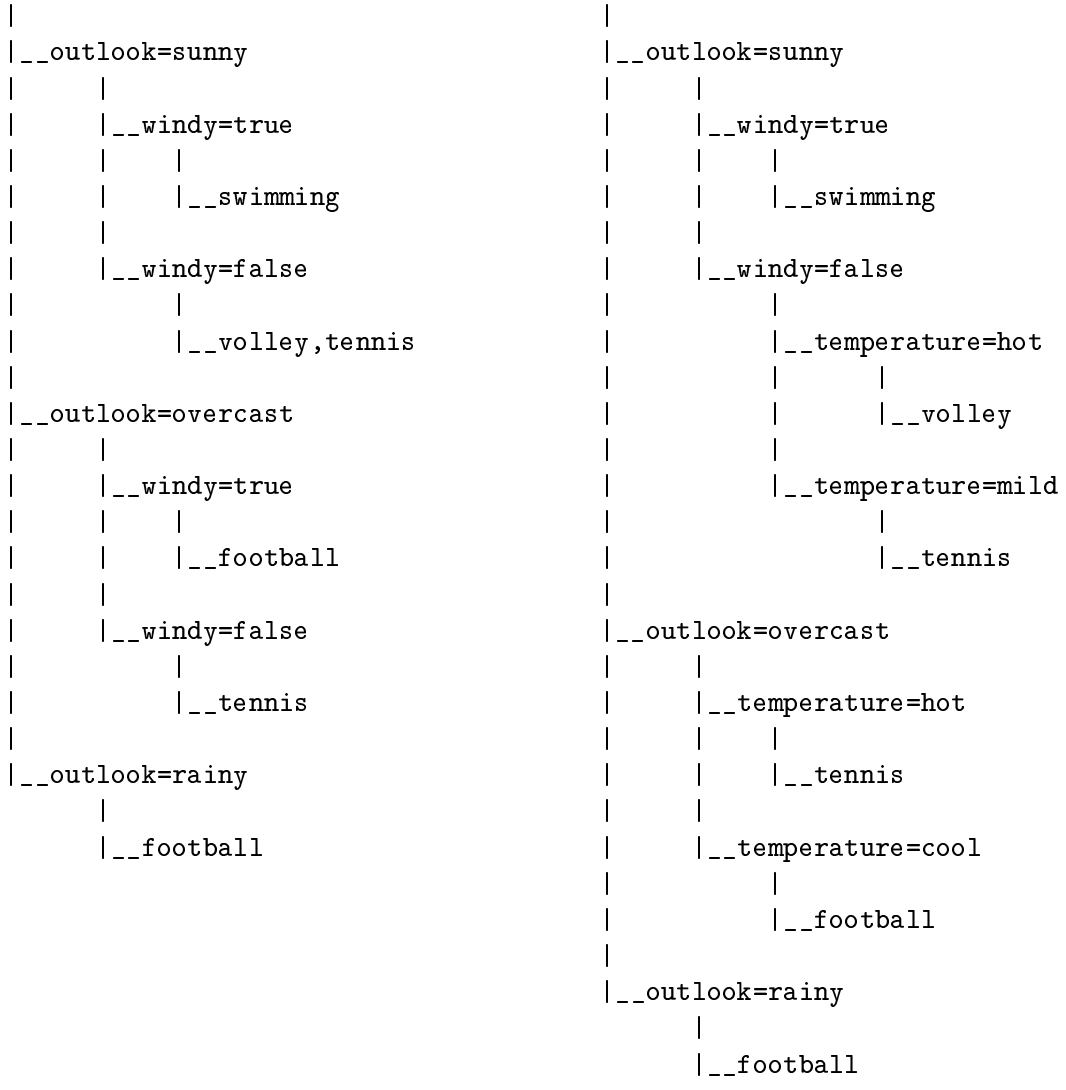
We select, as current branching attribute, the one which minimizes the conditioned observational entropy (which is equivalent to say that maximizes the observational information gain), and mark it as already used attribute.

Finally, for each newly generated child node the overall process should be iterated while all the following conditions hold:

1. There are remaining non used attributes.
2. The set of instances arriving to that node is not the empty set.
3. Observational entropy of current node is not below a predefined threshold value.

This procedure for building observational decision trees can be described according to the general framework detailed in section 4.3 as shown in *Observational Id3* column in table 4.5.

For the data in table 4.7 the induced observational decision tree and classical decision tree are depicted below:



4.5 A fuzzy sequential covering algorithm for the generation of rules.

In this section we present an algorithm which can be included in the sequential covering family of algorithms but adapted to accept vague information.

It's worth to recall that this kind of methods are intended to produce a disjunctive set of rules that describe a target concept by sequentially covering the set of instances representing this concept.

The method we are going to present follows this scheme in order to produce descriptions in terms of linguistic variables defined on the domains of the attributes.

4.5.1 Definition of the problem.

Let $A = \{A_1, \ldots, A_n\}$ be the set of attributes. We shall distinguish two relevant subsets: let $Z \subseteq A$ be the set of "explicative" attributes, that is, the attributes which will be used to generate the descriptions; and let $Y \subseteq A$ be the set of attributes which will provide intensional definitions for the concepts which are going to be described. $Z \cap Y = \emptyset$ is required in order to avoid tautological descriptions of concepts.[3]

Some attributes can be ignored in the process of generating descriptions by letting $Z \cup Y \neq A$.

Our approximation follows the one suggested by Yager [174] of partitioning the domains of the attributes and associating a template (a covering set of linguistic labels over a given domain) to each attribute in order to simplify the search space. For each attribute $A_i \in A$ we define a linguistic variable V_i.

We accept, as valid concepts, the elements of the boolean algebra generated by the terms of $\bigcup_{A_i \in Y} T(V_i)$.

Descriptions will be represented in disjunctive normal form, where literals will correspond to linguistic values of templates associated to the set Z of "explicative" attributes. In other words, descriptions are produced by the following grammar:

$$
\begin{aligned}
< description > &= < conjunctive\ rule > \vee \ldots \vee < conjunctive\ rule > \\
< conjunctive\ rule > &= < selector > \wedge \ldots \wedge < selector > \\
< selector > &= < template > = < linguistic\ value >
\end{aligned}
$$

Given that concepts and descriptions are described in terms of linguistic values of templates, and given that each linguistic value has its "meaning"

[3] For example, if we want to explain the concept "Temperature:High" in terms of attributes Temperature, Volume and Pressure, we want to avoid the rule: "Temperature:High if Temperature:High".

(fuzzy set) defined by the semantic rules of its associated linguistic variable, let us see how to combine these meanings in order to produce the meaning of descriptions and concepts.

We define this meaning in a constructive manner. Let $x =< a_1, \ldots, a_n >$ be an instance belonging to the set of instances X, (where $a_i \in Domain(A_i)$ is the value taken by x for attribute A_i), then:

1. Let S be the selector defined by the relationship $V_i = v$ (where v is a valid linguistic value of variable V_i). We define the meaning of S as:

$$\forall x \in X : \mu_S(x) = \mu_v(a_i) \tag{4.15}$$

   where μ_v is the meaning of the linguistic value v.

2. Let S_1, S_2 be two selectors. Then $\forall x \in X$ we define:

   - Meaning of "$S_1 \wedge S_2$":

$$\mu_{(S_1 \wedge S_2)}(x) = T(\mu_{S_1}(x), \mu_{S_2}(x)) \tag{4.16}$$

   - Meaning of "$S_1 \vee S_2$":

$$\mu_{(S_1 \vee S_2)}(x) = S(\mu_{S_1}(x), \mu_{S_2}(x)) \tag{4.17}$$

   where T is a t-norm and S a t-conorm.

Applying recursively these rules we should be able to construct the meaning of any description D or concept C.

4.5.2 Algorithm.

With the previous definitions, the problem could be stated as follows: given two concept definitions C^+ and C^-, find the description D which better "explains" C^+ and "excludes" C^- (intensional representation of counter-examples of the target concept C^+). We do not demand C^- be complementary of C^+.

Next let us present an algorithm to solve this task.

```
Generate Description Algorithm

    Input Parameters:
        PosExpr: target concept
        NegExpr: counter examples to exclude
        MinAccuracyThreshold: minimum accuracy for complexes
        MaxAccuracyThreshold: desired accuracy for complexes
```

```
    MinComplexCovering: minimum covering for complexes
    MaxDescriptionCovering: desired covering for descriptions
    BeamSize: size of the beam
Output Parameters:
    Description: output description

/* Initialize description */
Description = FALSE

/* Disjunctive generalization */
do
{
    /* Initialize new complex */
    Complex = TRUE
    ListOfComplexes.add(Complex);
    BestComplex = Complex

    /* Conjunctive specialization */
    while (BestComplex.accuracy < MaxAccuracyThreshold &&
           ListOfComplexes.size > 0)
    {
        /* Complexes specialization */
        Specialize(ListOfComplexes, ListOfSelectors)

        /* Updating BestComplex */
        ScoreComplexes(ListofComplexes, PosExpr, NegExpr)
        Purge(ListOfComplexes, MinComplexCovering, BeamSize)
        BestComplex = FindBestComplex(ListOfComplexes)
    }

    /* Description Generalization */
    if (BestComplex.accuracy >= MinAccuracyThreshold)
        GeneralizeDescription(Description, BestComplex)

} while (Description.covering > PastIterationDescription.covering &&
         Description.covering < MaxDescriptionCovering)

/* Output description */
return Description
```

Conjunctive specialization.

In this point we will present the specialization process for building complexes and the method for selecting the best complex according to a predefined criterion.

Given a current description D and a current beam B of previously generated complexes, for each complex C belonging to B, for all attribute $A_i \in Z$ and for all linguistic value (selector) v of template V_i associated to A_i, the following values are calculated (X being the set of instances):

- C^+ instances covered by the specialization $C \cap (V_i = v)$ and not already covered by description D:

$$\delta(C \cap (V_i = v), C^+, D) = \sum_{x \in X} T(\mu_C(x), \mu_v(x), \mu_{C^+}(x), N(\mu_D(x)))$$

- C^- instances covered by the specialization $C \cap (V_i = v)$ and not already covered by description D:

$$\delta(C \cap (V_i = v), C^-, D) = \sum_{x \in X} T(\mu_C(x), \mu_v(x), \mu_{C^-}(x), N(\mu_D(x)))$$

  where $\mu_v, \mu_{C^+}, \mu_{C^-}$ and μ_D are the meanings of v, C^+, C^- and D, respectively (calculated by applying 4.15, 4.16 and 4.17).

These values are used to choose the next selector which will specialize the current conjunctive rule C. The chosen linguistic value will be the one that maximizes the next function:

$$\sigma(C, v, C^+, C^-, D) = \frac{\delta(C \cap (V_i = v), C^+, D)}{\delta(C \cap (V_i = v), C^+, D) + \delta(C \cap (V_i = v), C^-, D)}$$

New generated complexes are then ranked by its value σ. Before, all those complexes with covering degree below `MinComplexCovering` threshold are purged.

Once ranked, only the first B (beam size) complexes are stored for iterating the specialization process. The rest are purged. The selection of parameter B should be a trading between the cost involved in the exploration of the search space and the risk inherent to greedy strategies.

The process of specialization finishes when the accuracy of the best complex C of the beam (computed as the degree of truth of the rule "If C then C^+) reaches a predefined threshold α. In other words, the next condition holds:

$$\frac{\sum_{x \in X} T(\mu_C(x), \mu_{C^+}(x))}{\sum_{x \in X} \mu_C(x)} > \alpha. \tag{4.18}$$

Disjunctive generalization.

At this point, a new conjunctive rule which covers a new "portion" of the set of instances representing the target concept is assumed to be generated. This new rule has to be incorporated into the current description, and then the process should be iterated.

Once a conjunctive rule C has been generated, the current description D is generalized by adding C as a new disjunctive term.

This process of generation of conjunctive rules and generalization of the current description iterates until the degree of covering of description D (calculated as the degree of truth of the rule: "If C^+ then D") reaches a threshold β:

$$\frac{\sum_{x \in X} T(\mu_D(x), \mu_{C^+}(x))}{\sum_{x \in X} \mu_{C^+}(x)} > \beta. \tag{4.19}$$

It should be noted the difference between covering and accuracy: the former measures the necessary condition of the bi-equivalence $D \Leftrightarrow C^+$, while the latter deals with the sufficient condition.

Finally, when condition 4.19 holds, description D is returned as the result.

4.5.3 Experimental results.

Introduction.

According to what was already stated, this chapter is mainly concerned to devise learning methods whose induced descriptions are expected to be comprehensible. Henceforth, the use of linguistic values as the representation language was entirely advisable.

Notwithstanding this emphasis on understandability, accuracy should not be left aside. This section is devoted to present accuracy comparisons between the fuzzy sequential covering algorithm (FSQ) introduced in the previous section and a selected set of classical machine learning algorithms. More precisely, algorithms will be compared according to their attained results in supervised classification problems.

FSQ induces a set of rules accounting for the description of a given target. This behavior can easily be adapted to classification problems in which characterization rules intended to predict the proper class for unseen instances are searched . To achieve this goal, FSQ is run as many times as

classes defined for the class attribute. Let us suppose a class attribute C with three possible classes $\{C_1, C_2, C_3\}$. FSQ is run three times with the following configurations:

1. $C^+ = C_1$, $C^- = C_2 \vee C_3$
2. $C^+ = C_2$, $C^- = C_1 \vee C_3$
3. $C^+ = C_3$, $C^- = C_1 \vee C_2$

A set of rules R_i is induced for each class C_i. The procedure for classifying a new instance x consists of computing the degree of truth of the proposition "$x \in R_i$" for each rule set R_i. Finally, the class whose associated description has maximum degree of truth is selected as the output class.

Methodology.

Experiments will be performed assuming a 10-fold cross validation scheme, where data sets are randomly divided into ten disjoints subsets, each containing approximately the same number of instances. Then, for each partition, the remaining nine partitions are used as the training set to induce a new classifier while the current partition itself serves as test data. After training and testing, an estimate of the classifier error rate is obtained. The ten cross validations estimates are then averaged to provide a global estimate for accuracy. Finally, algorithms are compared determining the level of significance that one algorithm outperforms another. Classification error rates and standard deviations are considered to differ from one another significatively if the level of a paired t-test is below 0,05.

Data sets.

Data sets from the public repository of the University of Irvine [121] have become a standard benchmark for the machine learning community. The UCI Irvine archive is a repository of data sets encompassing a wide variety of data types, analysis tasks and application areas. The primary role of this repository is to enable researchers to perform empirical analysis of machine learning algorithms.

In order to cover the whole range, we have selected a subset of data sets having the set of predictive attributes all nominal, all numeric, and a mixture of numeric and nominal attributes. Since we are interested in comparing the classification accuracy of algorithm FSQ with respect to a representative set of well known standard algorithms, data sets with non nominal class attribute will not be considered.

The selected data sets are summarized in the table below:

Dataset	Predictive attributes	Numeric	Nominal
Breast Cancer (BCW)	9	9	0
Hepatitis (HP)	19	6	13
Iris (IR)	4	4	0
Lenses (LS)	4	0	4
Mushroom (MH)	22	0	22
Weather (WH)	4	0	4
Wine (WN)	13	13	0
Zoo (ZO)	17	1	16

Algorithms.

We have chosen a set of representative classifiers as the bench mark for performing the experiments. Their accuracies will be compared to that of FSQ in order to draw conclusions regarding the classification abilities of the algorithms involved.

The set of selected algorithms is listed below.

- Conjunctive Rule (CR): single conjunctive rule learner. A rule consists of a conjunction of antecedents and the consequent (class value). If the test instance is not covered by the induced rule, then it is predicted using the default class (majority class) of data not covered by the rule.

- Decision Table (DT): algorithm for building a simple decision table majority classifier [98].

- Nearest Neighbor (NN): nearest neighbor-like algorithm using non nested generalized exemplars [111].

- One Rule (OR): simple classifier using just the minimum error attribute as predictor for class attribute. Numeric attributes are discretized [73].

- Part (PART): algorithm for generating a PART decision list by building a partial C4.5 decision tree in each iteration and making the best leaf into a rule.

- Ripple-down Rule Learner (RDL): generates a default rule first, and then the exceptions for the default rule with the least error rate. Then it generates the "best" exceptions and iterates until pure. A tree-like expansion of exceptions is performed where exceptions are a set of rules that predict classes other than the default.

- Majority (MA): classifier predicting the mean or the mode, for numeric and nominal classes respectively.
- Propositional Rules (PRL): algorithm implementing a propositional rule learner as described in [23].
- Decision Stump (DS): algorithm for building and using a decision stump.
- Decision Tree (J48): algorithm for generating C4.5 decision trees [135].
- Logistic Model Tree (LMT): procedure for building logistic model trees which are classification trees with logistic regression functions at the leaves.
- Random Tree (RT): algorithm for constructing a tree that considers k randomly chosen attributes at each node.
- Fast Decision Tree (FDT): Fast decision tree learner.
- Random Forest (RF): algorithm which builds a forest of random trees [15].
- Prism (PRISM): algorithm for building and using a PRISM rule set for classification [19].
- Decision Tree (ID3): classical algorithm for induction of decision trees [133].

Table 4.11 shows whether a given algorithm can handle only numerical, only nominal or a mixture of numerical and nominal predictive attributes.

Results.

In this section we will summarize classification accuracies for each data set, and the result of their statistical comparison according to the methodology already described.

Tables 4.13 and 4.12 show mean and standard deviation averaged over 10 cross validation runs, for each data set and algorithm considered.

Based on these results, a paired t-test is performed to determine whether one algorithm significatively outperforms another.

Table 4.14 describes, for each data set, whether FSQ algorithm performs better, worse or "equal" [4]. When FSQ performs better than algorithm A, the sign "-" is depicted. On the contrary, if FSQ performs worse, sign "+" is represented. Sign "=" appears when no conclusions can be significatively drawn.

[4]Term "equal" refers to the situation when no better or worse performance can be inferred, given a predefined confidence degree.

Algorithm	**Numerical attributes**	**Nominal attributes**
CR	yes	yes
DT	yes	yes
NN	yes	yes
OR	yes	yes
PART	yes	yes
RDL	yes	yes
MA	yes	yes
PRL	yes	yes
DS	yes	yes
J48	yes	yes
LMT	yes	yes
RT	yes	yes
FDT	yes	yes
RF	yes	yes
PRISM	no	yes
ID3	no	yes

Table 4.11: Algorithms and types of attributes.

Table 4.12: Accuracies and standard deviations (I).

Set	**FSQ**	**CR**	**DT**	**NN**	**OR**	**PART**	**RDL**	**PRL**
BCW	93.54 (1.61)	91.35 (3.36)	95.9 (2.84)	96.49 (1.85)	91.79 (3.14)	95.45 (3.14)	95.75 (2.01)	96.49 (1.85)
HP	83.19 (3.52)	83.75 (6.04)	86.25 (12.43)	87.5 (8.33)	77.5 (7.91)	83.75 (15.65)	85 (9.86)	83.75 (10.29)
IR	96.84 (4.08)	66.67 (0)	93.33 (5.44)	96 (4.66)	94 (5.84)	94 (5.84)	94 (6.63)	94 (5.84)
WN	89.04 (2.27)	63.5 (6.48)	91.63 (7.04)	97.75 (2.91)	76.41 (8.98)	92.71 (5.33)	93.86 (4.86)	91.57 (9.27)
ZO	93.2 (6.55)	59.55 (5.86)	91.18 (7.09)	96.18 (6.54)	73.27 (10.54)	92.18 (8.94)	88.27 (8.62)	90.27 (9.99)
LS	62.17 (30.42)	60.33 (25.71)	78.83 (24.83)	75.5 (27.97)	72.17 (27.12)	83.5 (22.54)	86.33 (21.89)	80 (25.84)
MH	98.2 (0.96)	89.86 (1.27)	100 (0)	100 (0)	98.44 (0.61)	100 (0)	100 (0)	100 (0)
WH	55 (43.78)	60 (39.44)	45 (43.78)	85 (24.15)	35 (41.16)	60 (45.95)	40 (45.95)	70 (42.16)

Table 4.13: Accuracies and standard deviations (II).

Set	**MA**	**DS**	**J48**	**LMT**	**RF**	**RT**	**FDT**	**PR**	**ID3**
BCW	65.01 (0.48)	91.94 (3.62)	96.05 (2.4)	96.63 (2.3)	96.63 (2.5)	94.28 (3.21)	95.31 (2.47)	na na	na na
HP	83.75 (6.04)	83.75 (6.04)	86.25 (9.22)	85 (14.19)	86.25 (7.1)	90 (9.86)	78.75 (13.24)	na na	na na
IR	33.33 (0)	66.67 (0)	96 (5.62)	94 (4.92)	95.33 (5.49)	90.67 (10.52)	94.67 (5.26)	na na	na na
WN	39.93 (2.6)	57.75 (6.51)	93.3 (5.1)	97.22 (5.4)	98.3 (2.74)	93.2 (6.48)	94.41 (5.24)	na na	na na
ZO	40.64 (3.48)	60.45 (3.77)	92.18 (8.94)	95.18 (8.15)	96.09 (5.05)	89.27 (9.39)	90.27 (9.99)	na na	na na
LS	64.33 (23.69)	72.17 (27.12)	83.5 (22.54)	na na	76 (28.26)	61.33 (32.21)	72.5 (28.46)	65.17 (28.34)	73.17 (29.95)
MH	61.8 (0.07)	89.86 (1.27)	100 (0)	na na	100 (0)	99.91 (0.13)	100 (0)	100 (0)	100 (0)

(to be continued)

									(continuation)
WH	70 (34.96)	30 (34.96)	55 (43.78)	na na	70 (42.16)	75 (35.36)	70 (34.96)	85 (33.75)	85 (33.75)

Table 4.14: Algorithms comparison (I).

Set	**CR**	**DT**	**NN**	**OR**	**PART**	**RDL**	**PRL**
BCW	=	=	=	=	=	=	=
HP	=	=	=	=	=	=	=
IR	-	=	=	=	=	=	=
WN	-	=	+	=	=	=	=
ZO	-	=	=	-	=	=	=
LS	=	+	=	=	+	+	+
MH	-	+	+	=	+	+	+
WH	=	=	=	=	=	=	=

Table 4.15: Algorithms comparison (II).

Set	**MA**	**DS**	**J48**	**LMT**	**RF**	**RT**	**FDT**	**PR**	**ID3**
BCW	-	=	=	=	=	=	=	na	na
HP	=	=	=	=	=	=	=	na	na
IR	-	-	=	=	=	=	=	na	na
WN	-	-	=	=	+	=	=	na	na
ZO	-	-	=	=	=	=	=	na	na
LS	=	=	+	na	=	=	=	=	=
MH	-	-	+	na	+	+	+	+	+
WH	=	=	=	na	=	=	=	=	=

Conclusions.

This chapter has been concerned to the study and development of methods intended to produce comprehensible descriptions of the relationships existing in the data, involving at the same time, a proper management of uncertainty.

Consequently, emphasis has not solely focused on achieving high accuracy. The question now is if this bias towards comprehensibility, together

with the management of uncertainty through the use of linguistic labels, affect significatively the resulting accuracy when compared to standard learning algorithms. The presented results show that performance of FSQ algorithm is comparable with that of standard algorithms.

Regarding the achievement of comprehensibility, for illustrative purposes let us analyze the descriptions obtained by FSQ from the Monk's data. These data sets have the particularity of having been generated according to predefined and known rules. We can measure the quality and compactness of the obtained descriptions by comparing them to the set of "seed" rules.

The three data sets composing the Monks problems are based on six nominal attributes $\{a_1, \ldots, a_6\}$ which can take values over the set $\{1, 2, 3, 4\}$, and a binary class attribute.

Each problem involves learning a binary function defined over this domain. The target concepts associated to the Monks problems are:

- Monks first problem: $(a_1 = a_2)$ or $(a_5 = 1)$.
- Monks second problem: exactly two of $(a_1 = 1, a_2 = 1, a_3 = 1, a_4 = 1, a_5 = 1, a_6 = 1)$.
- Monks third problem: $(a_5 = 3$ and $a_4 = 1)$ or $(a_5 \neq 4$ and $a_2 \neq 3)$.

First problem is in standard disjunctive normal form and is supposed to be easy learnable by algorithms capable of producing DNF descriptions. Conversely, second problem is similar to parity problems where attributes are combined in such a way that makes it complicated to describe in DNF using the given attributes only. Third problem is again in DNF.

The descriptions obtained by FSQ for first and third problems are:

- Monks first problem:

```
a5:1

    OR

a1:1 AND a2:1

    OR

a1:2 AND a2:2

    OR

a1:3 AND a2:3
```

- Monks third problem:

```
a5:3 AND a4:1
    OR
a2:1 AND a5:1
    OR
a2:1 AND a5:2
    OR
a2:1 AND a5:3
    OR
a2:2 AND a5:1
    OR
a2:2 AND a5:2
    OR
a2:2 AND a5:3
    OR
a2:4 AND a5:1
    OR
a2:4 AND a5:2
    OR
a2:4 AND a5:3
```

which clearly correspond to DNF expressions for the above seeding rules.

As expected, the second monk's problem could not be learnable by FSQ since it can not be represented in DNF form.

These results show that FSQ has captured quite well the set of "seed" rules in any case with the exception of the second problem.

4.6 Generating indistinguishability operators from prototypes.

In many situations, given an indistinguishability operator E on a universe of discourse X, it is useful to define an indistinguishability operator on a set of fuzzy subsets of X compatible with E. This is the case in approximate reasoning or in fuzzy control [43, 91] and there are some standard ways to generate it [10]. Nevertheless, the opposite problem has not been studied deeply since now, although it seems a very interesting one.

To focus on the problem, let us consider the following situation: let us suppose the existence of some prototypes $a_1, a_2, ..., a_n$, an indistinguishability operator $\bar{E}$ between them and a set X of objects resembling the prototypes to some extent. Then it seems reasonable to extend the relation $\bar{E}$ to the set X.

The preceding situation can be modelled in this way: there are n fuzzy subsets of X denoting the resemblance of the elements of X to the prototypes and an indistinguishability operator $\bar{E}$ between these prototypes. The question is how to define an indistinguishability operator E on X compatible with $\bar{E}$.

This section studies a couple of ways to generate such an indistinguishability operator related to the ones used in approximate reasoning and a third one based on the duality principle [10, 124]. Two interesting cases are when the fuzzy subsets of X define a partition or a hard-partition on X.

4.6.1 An "optimistic" method.

Given an indistinguishability operator E on a universe of discourse X, there are some standard ways to generate indistinguishability operators on some fuzzy subsets of X compatible with E.

Proposition 4.6.1 *Given a set P of fuzzy subsets of a set X, the fuzzy relation E^* on P defined $\forall \mu, \nu \in P$ by:*

$$E^*(\mu, \nu) = \inf_{x \in X} \overleftrightarrow{T}(\mu(x), \nu(x)) \tag{4.20}$$

is a T-indistinguishability operator on P.

Proposition (4.6.1) allows us to generate an indistinguishability operator on P when in the universe of discourse X the trivial equality relation is assumed. If there is another indistinguishability operator E defined on X, the

following proposition allows us to generate an indistinguishability operator on P compatible with E.

Proposition 4.6.2 *Given a set P of fuzzy subsets of a set X and a T-indistinguishability operator E on X, the fuzzy relation $\bar{E}$ on P defined by*

$$\bar{E}(\mu,\nu) = \inf_{x\in X} \overset{\leftrightarrow}{T}(\phi_E(\mu)(x), \phi_E(\nu)(x)) \tag{4.21}$$

is a T-indistinguishability operator , where ϕ_E is the upper approximation with respect to E (definition (1.5.9)*).*

Therefore, the degree of similarity between μ and ν via $\bar{E}$ is the degree of similarity between their upper approximations by observable sets with respect to E.

Proposition 4.6.3 *If P contains the set of columns of E, then there is an isometric embedding of X into P:*

$$E(x,y) = E^*(\phi_E(\{x\}), \phi_E(\{y\}))). \tag{4.22}$$

Let us now focus to the problem of defining an indistinguishability operator E on a set X compatible with an indistinguishability operator defined between some fuzzy subsets of X.

In order to clarify the problem, let us first consider the crisp case: Given an equivalence relation $\sim_P$ defined on a subset P of the power set of a universe X, an equivalence relation $\sim$ between the elements of X compatible with $\sim_P$ is searched.

A first attempt is to define $\sim$ by $x \sim y$ if and only if there exist A, B of P such that $x \in A \wedge y \in B \wedge A \sim_P B$.

Nevertheless, this definition does not give in general an equivalence relation. For instance, if $X = \{x, y, z, t\}$, $P = \{A, B, C, D\}$ with $A = \{x\}$, $B = \{y\}$, $C = \{y, t\}$, $D = \{z\}$ and $\sim_P$ is the equivalence relation on P that partitions P in the two equivalence classes $\{A, B\}$ and $\{C, D\}$, then $x \sim y$, $y \sim z$ but x is not related to z.

In order to obtain an equivalence relation we must add a compatibility condition such as

Definition 4.6.4 *Crisp compatibility condition: If $x \in A$ and $x \in B$ and $A \sim_P C$, then $B \sim_P C$.*

The preceding study leads to the following definition and compatibility relation in the fuzzy framework:

Definition 4.6.5 *Given a set P of fuzzy subsets of a set X and a T-indistinguishability operator $\bar{E}$ on P, the fuzzy relation E_4 on X is defined by*

$$E_4(x,y) = \sup_{\mu,\nu\in P} T(\mu(x),\nu(y),\bar{E}(\mu,\nu))\ \forall x,y\in X. \tag{4.23}$$

As in the crisp case, E_4 is not a T-indistinguishability operator in general and we need to add a compatibility condition:

Definition 4.6.6 *Compatibility condition: Given a set P of fuzzy subsets of a set X and a T-indistinguishability operator $\bar{E}$ on P, we say that $\bar{E}$ satisfies the compatibility condition if and only if $\forall \mu,\nu,\rho\in P$:*

$$T(\mu(x),\nu(x),\bar{E}(\mu,\rho)) \leqslant \bar{E}(\nu,\rho)\ \forall x,y\in X. \tag{4.24}$$

It is worth noticing that if P is a hard-partition of X, then this condition is trivially satisfied.

Proposition 4.6.7 *Assuming the compatibility condition of definition* (4.6.6)*, the fuzzy relation E_4 defined in* (4.6.5) *is symmetric, T-transitive but not necessarily reflexive.*

Proof 4.6.8

$$(*)\, T(E_4(x,y),E_4(y,z)) =$$
$$T(\sup_{\mu,\nu\in P} T(\mu(x),\nu(y),\bar{E}(\mu,\nu)),\ \sup_{\rho,\tau\in P} T(\rho(y),\tau(z),\bar{E}(\rho,\tau))).$$

Fixing $\mu,\nu,\rho,\tau\in P$ and by condition (4.6.6)

$$(**)\, T(\mu(x),\nu(y),\bar{E}(\mu,\nu),\rho(y),\tau(z),\bar{E}(\rho,\tau)) \leqslant$$

by transitivity of $\bar{E}$

$$T(\mu(x),\tau(z),\bar{E}(\mu,\rho),\bar{E}(\rho,\tau)) \leqslant$$
$$(***)\, T(\mu(x),\tau(z),\bar{E}(\mu,\tau)) \leqslant$$
$$\sup_{\sigma,\pi\in P} T(\sigma(x),\pi(z),\bar{E}(\sigma,\pi)) =$$
$$E_4(x,z)$$

Since inequality (**) $\leq$ (***) holds $\forall \mu,\nu,\rho,\tau\in P$, we get (*) $\leq$ (***).
□

Lemma 4.6.9 *If $\forall x\in X$ there exists $\mu\in P$ such that $\mu(x)=1$, then E_4 is reflexive (and therefore a T-indistinguishability operator).*

Proof 4.6.10 *Trivial: Given $x \in X$, let $\mu \in P$ be such that $\mu(x) = 1$. Then $E_4(x,x) = T(\mu(x), \mu(x), \bar{E}(\mu,\mu)) = 1$.* $\square$

Lemma 4.6.11 *If P is a finite set, then the reciprocal of the previous lemma holds.*

Proof 4.6.12 *If P is finite, then there exist μ, ν of P such that $T(\mu(x), \nu(x), \bar{E}(\mu,\nu)) = 1$ and therefore $\mu(x) = 1$.* $\square$

An interesting question is whether the elements of P are observable with respect to E. The following proposition gives a sufficient condition for this property to be satisfied:

The following result will be needed in proposition (4.6.14):

Proposition 4.6.13 *[10] Given a set P of fuzzy subsets of a set X, a T-indistinguishability operator E on X and an element x of X, the fuzzy subset x^{**} of P defined by*

$$x^{**}(\mu) = \mu(x) \tag{4.25}$$

is extensional with respect to E^*.

Proposition 4.6.14 *Assuming condition* (4.6.6)*, if $\bar{E} \leqslant E^*$ then the elements of P are observable sets with respect to E_4.*

Proof 4.6.15 *Given x, y of X,*

$$\begin{aligned} T(E,x,y),\mu(x)) &= T(\sup_{\nu,\rho\in P} T(\nu(x),\rho(y),\bar{E}(\nu,\rho)),\mu(x)) \\ &= \sup_{\nu,\rho\in P} T(\nu(x),\rho(y),\mu(x),\bar{E}(\nu,\rho)) \\ &\leqslant \sup_{\rho\in P} T(\rho(y),\bar{E}(\mu,\rho)) \\ &\leqslant \sup_{\rho\in P} T(\rho(y),E^*(\mu,\rho)) \\ &\leqslant \mu(y). \end{aligned}$$

The last inequality follows from proposition (4.6.13) *since y^{**} is extensional with respect to E^*.* $\square$

At this point, we have a way to define an indistinguishability operator E^* (proposition (4.6.2)) on P starting on from an operator E defined on X, and reciprocally. The following result shows that if from an E on X we generate E^* on P and then from this E^* we generate E_4^* on X, then $E_4^* = E$ if P is the set of columns of E.

Proposition 4.6.16 *Let E be a T-indistinguishability operator on a set X and P the set of columns of E. If from E^* we define a fuzzy relation E_4^* on X as in definition* (4.6.5)*, then $E_4^* = E$.*

Proof 4.6.17 *By proposition* (4.6.3)

$$\begin{aligned} E^*(\phi_E(x), \phi_E(y)) &= E(x,y) \\ E_4^*(x,y) &= \sup_{z,t \in X} T(E(x,z), E(y,t), E^*(\phi_E(z), \phi_E(t))) \\ &= \sup_{z,t \in X} T(E(x,z), E(y,t), E(z,t)) \\ &\leqslant E(x,y) \end{aligned}$$

where last inequality follows from the transitivity of E.
Taking $z = x$ and $t = y$ equality follows. $\square$

4.6.2 A "conservative" method.

Definition (4.6.5) allows the generation of a (non necessarily reflexive) indistinguishability operator E on X when P is a hard-partition, since in this case condition of definition (4.6.6) is fulfilled trivially. Since reflexivity is not a very important condition in this context, this gives a good tool to work with when starting on from a hard-partition.

In this Section, we give another way to generate an indistinguishability operator E on X compatible with an indistinguishability operator defined on a set P of fuzzy subsets of X that works when P is a partition.

Definition 4.6.18 *Given a set P of fuzzy subsets of a set X and a T-indistinguishability operator $\bar{E}$ on P, the fuzzy relation E_5 on X is defined by*

$$E_5(x,y) = \inf_{\mu,\nu \in P} \hat{T}(T(\mu(x), \nu(y)) | \bar{E}(\mu,\nu)) \ \forall x, y \in X. \tag{4.26}$$

In the crisp case, this means that

$$x \sim y \text{ if and only if } \forall \mu, \nu \in P : \ (x \in \mu \wedge y \in \nu) \Rightarrow \mu \sim_P \nu \tag{4.27}$$

Lemma 4.6.19 *E_5 is reflexive if and only if for all μ, ν of P*

$$T(\mu(x), \nu(x)) \leqslant \bar{E}(\mu,\nu). \tag{4.28}$$

Proof 4.6.20

$$E_5(x,x) = \inf_{\mu,\nu\in P} \hat{T}(T(\mu(x),\nu(x)|\bar{E}(\mu,\nu)) = 1$$

if and only if $\forall \mu,\nu \in P$*:*

$$T(\mu(x),\nu(x)) \leqslant \bar{E}(\mu,\nu).$$

The condition of the previous lemma is trivially satisfied if P is a hard-partition of X.

Lemma 4.6.21 *E_5 is symmetric.*

In order to study the transitivity of E_5 we need the following lemma:

Lemma 4.6.22 $\forall x,y,z,t \in [0,1]$:

$$T(\hat{T}(x|y),\hat{T}(z|t)) \leqslant \hat{T}(T(x,z)|T(y,t)). \tag{4.29}$$

Proof 4.6.23 *Given*

$$A = \{\alpha \in [0,1]|T(\alpha,x) \leqslant y\}$$

and

$$B = \{\beta \in [0,1]|T(\beta,z) \leqslant t\}$$

and given $\alpha \in A$ and $\beta \in B$,

$$T(\alpha,x,\beta,z) \leqslant T(y,t).$$

Then

$$\begin{aligned}\hat{T}(T(x,z)|T(y,t)) = \sup\{\gamma \in [0,1]|T(x,z,\gamma) \\ \leqslant T(y,t)\}.\end{aligned}$$

Therefore $\forall \alpha \in A$ and $\forall \beta \in B$

$$T(\alpha,x,\beta,z) \leqslant T(y,t)$$

and therefore

$$T(\sup A, \sup B) \leqslant T(y,t). \qquad \square$$

Proposition 4.6.24 *If for all x of X there exists μ of P such that $\mu(x) = 1$, then E_5 is T-transitive.*[5]

Proof 4.6.25

$$T(E_5(x,y), E_5(y,z)) =$$
$$= T(\inf_{\mu,\nu\in P} \hat{T}(T(\mu(x),\nu(y))|\bar{E}(\mu,\nu)), \inf_{\rho,\tau\in P} \hat{T}(T(\rho(y),\tau(z))|\bar{E}(\rho,\tau)))$$
$$= \inf_{\mu,\nu,\rho,\tau\in P} T(\hat{T}(T(\mu(x),\nu(y))|\bar{E}(\mu,\nu)), \hat{T}(T(\rho(y),\tau(z))|\bar{E}(\rho,\tau)))$$

taking $\rho = \nu$

$$\leqslant \inf_{\mu,\nu,\tau\in P} T(\hat{T}(T(\mu(x),\nu(y))|\bar{E}(\mu,\nu)), \hat{T}(T(\nu(y),\tau(z))|\bar{E}(\nu,\tau)))$$

by lemma (4.6.22)

$$\leqslant \inf_{\mu,\nu,\tau\in P} \hat{T}(T(\mu(x),\nu(y),\nu(y),\tau(z))|T(\bar{E}(\mu,\nu),\bar{E}(\nu,\tau)))$$

$\hat{T}$ is non decreasing in the second variable and $\bar{E}$ is transitive

$$\leqslant \inf_{\mu,\nu,\tau\in P} \hat{T}(T(\mu(x),\nu(y),\nu(y),\tau(z))|\bar{E}(\mu,\tau))$$

taking ν with $\nu(y) = 1$

$$\leqslant \inf_{\mu,\tau\in P} \hat{T}(T(\mu(x),\tau(z))|\bar{E}(\mu,\tau))$$
$$= E_5(x,z). \qquad \square$$

4.6.3 A method based on the duality principle.

The duality principle [10, 124] allow us to consider the elements of a universe X as fuzzy subsets and gives a way to make operate these elements over fuzzy subsets of X.

Following this idea, we will generate an indistinguishability operator on X when such an operator is defined between some of their fuzzy subsets in a very natural way.

Definition 4.6.26 *[10] Let X be a set and P a set of fuzzy subsets of X. Given $x \in X$, the fuzzy subset x^{**} of P defined $\forall \mu \in P$ as*

$$x^{**}(\mu) = \mu(x) \tag{4.30}$$

is called the dual of x.

[5] The condition of this proposition is satisfied if P is a partition of X.

Definition 4.6.27 *Let P be a set of fuzzy subsets of X. P is said to separate points if and only if $\forall x, y \in X\ \exists \mu \in P$ such that $\mu(x) \neq \mu(y)$.*

Proposition 4.6.28 *If P separates points, then the map that assigns to every element x of X its dual x^{**} is a bijection.*

Proposition 4.6.29 *Given a set P of fuzzy subsets of X and $\bar{E}$ a T-indistinguishability operator on P, the fuzzy relation $\bar{E}^*$ on the set X^{**} of dual elements of X defined for all $x^{**}, y^{**} \in X^{**}$ by*

$$\bar{E}^*(x^{**}, y^{**}) = \inf_{\mu \in P} \overleftrightarrow{T} (\phi_{\bar{E}}(x^{**})(\mu), \phi_{\bar{E}}(y^{**})(\mu)) \tag{4.31}$$

is a T-indistinguishability operator .

So the degree of similarity between x^{**} and y^{**} is the degree of similarity between their respective upper approximations $\phi_{\bar{E}}(x^{**})$ and $\phi_{\bar{E}}(y^{**})$ by observable sets with respect to $\bar{E}$.

Definition 4.6.30 *Let P be a set of fuzzy subsets of a set X and $\bar{E}$ a T-indistinguishability operator on P. The T-indistinguishability operator E_6 on X is defined by*

$$E_6(x, y) = \bar{E}^*(x^{**}, y^{**})\ \forall x, y \in X. \tag{4.32}$$

Proposition 4.6.31 *Let E be a T-indistinguishability operator on a set X, P a set of fuzzy subsets of X and $\bar{E}$ the T-indistinguishability operator on P defined for all $\mu, \nu \in P$ by*

$$\bar{E}(\mu, \nu) = \inf_{x \in X} \overleftrightarrow{T} (\phi_E(\mu)(x), \phi_E(\nu)(x)). \tag{4.33}$$

If E_6 is the T-indistinguishability operator generated from $\bar{E}$ using definition (4.6.30)*, then $E_6 = E$ if and only if P is a generating family of E.*

Proof 4.6.32

$$E_6(x, y) = \inf_{\mu \in P} \overleftrightarrow{T} (\phi_{\bar{E}}(x^{**})(\mu), \phi_{\bar{E}}(y^{**})(\mu))$$

by proposition (4.6.13)

$$= \inf_{\mu \in P} \overleftrightarrow{T} ((x^{**})(\mu), (y^{**})(\mu))$$

$$= \inf_{\mu \in P} \overleftrightarrow{T} (\mu(x), \mu(y))$$

□

which is equal to $E(x, y)$ if and only if P is a generating family of E.

4.6.4 An example.

In the following example we will generate the T-indistinguishability operators E_4, E_5 and E_6 from a given T-indistinguishability operator $\bar{E}$.

A video club has its videos classified in labels; among them: 'Drama' (D), 'Adventures' (A), 'Western' (W), 'Comedy' (Co) and 'Children' (Ch). The owner would like to have some way to recommend a second film to a client when he or she withdraws a video. For doing so, he passed a questionnaire to his clients asking them to tick the types of films they like and realized that, for example, 85% of the clients that had ticked 'Western' or 'Adventures' also had ticked the other one , and obtaining the following matrix $\bar{E}$ (see [49]) that corresponds to a T-indistinguishability operator (being T the Lukasiewicz t-norm).

$$\begin{array}{c} \\ D \\ A \\ W \\ Co \\ Ch \end{array} \begin{array}{c} \begin{array}{ccccc} D & A & W & Co & Ch \end{array} \\ \left(\begin{array}{ccccc} 1 & 0.3 & 0.3 & 0.1 & 0 \\ 0.3 & 1 & 0.85 & 0.5 & 0.5 \\ 0.3 & 0.85 & 1 & 0.3 & 0.4 \\ 0.1 & 0.5 & 0.3 & 1 & 0.5 \\ 0 & 0.5 & 0.4 & 0.5 & 1 \end{array} \right) \end{array} = \bar{E}$$

He also realized that many films could have more than one label, at least at some extent. For example, 'Gone with the wind' is labelled as 'Drama' but at some extent is also a Western. Assigning 0, 0.25, 0.5, 0.75 or 1 depending on the degree in which he thinks a film could have the corresponding label he obtains the following matrix M for the films 'Gone with the wind' (GW), 'Casablanca' (CB), 'The Rush of Gold' (RG), 'Indiana Jones' (I) and 'High Noon' (H).

$$\begin{array}{c} \\ D \\ A \\ W \\ Co \\ Ch \end{array} \begin{array}{c} \begin{array}{ccccc} GW & CB & RG & I & H \end{array} \\ \left(\begin{array}{ccccc} 1 & 1 & 0.25 & 0 & 0.25 \\ 0.25 & 0.25 & 0.25 & 1 & 0 \\ 0.5 & 0 & 0.25 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0.75 & 0.25 & 0 \end{array} \right) \end{array} = M$$

The files of the matrix give the labels as fuzzy subsets of the set of videos while the columns are the dual fuzzy subsets of the videos acting on the set of labels.

From $\bar{E}$ and M we obtain the T-indistinguishability operators E_4, E_5 and E_6 with matrices:

$$
\begin{array}{c} \\ GW \\ CB \\ RG \\ I \\ H \end{array}
\begin{array}{c}
\begin{array}{ccccc} GW & CB & RG & I & H \end{array} \\
\left(\begin{array}{ccccc}
1 & 1 & 0.1 & 0.3 & 0.3 \\
1 & 1 & 0.1 & 0.3 & 0.3 \\
0.1 & 0.1 & 1 & 0.5 & 0.4 \\
0.3 & 0.3 & 0.5 & 1 & 0.85 \\
0.3 & 0.3 & 0.4 & 0.85 & 1
\end{array} \right)
\end{array} = E_4
$$

$$
\begin{array}{c} \\ GW \\ CB \\ RG \\ I \\ H \end{array}
\begin{array}{c}
\begin{array}{ccccc} GW & CB & RG & I & H \end{array} \\
\left(\begin{array}{ccccc}
0.8 & 0.8 & 0.1 & 0.3 & 0.3 \\
0.8 & 1 & 0.1 & 0.3 & 0.3 \\
0.1 & 0.1 & 0.75 & 0.5 & 0.4 \\
0.3 & 0.3 & 0.5 & 1 & 0.85 \\
0.3 & 0.3 & 0.4 & 0.85 & 1
\end{array} \right)
\end{array} = E_5
$$

$$
\begin{array}{c} \\ GW \\ CB \\ RG \\ I \\ H \end{array}
\begin{array}{c}
\begin{array}{ccccc} GW & CB & RG & I & H \end{array} \\
\left(\begin{array}{ccccc}
1 & 0.8 & 0.1 & 0.3 & 0.3 \\
0.8 & 1 & 0.1 & 0.3 & 0.3 \\
0.1 & 0.1 & 1 & 0.5 & 0.4 \\
0.3 & 0.3 & 0.5 & 1 & 0.85 \\
0.3 & 0.3 & 0.4 & 0.85 & 1
\end{array} \right)
\end{array} = E_6
$$

According to any of the three matrices he would suggest 'Casablanca' to a client withdrawing 'Gone with the wind' and vice versa. The same occurs with the couple 'Indiana' and 'High Noon', and to a client withdrawing 'The Gold Rush' he would recommend 'Indiana' or 'High Noon'.

Chapter 5

Summary of contributions and future work.

This dissertation is a contribution to the study of uncertainty from the notion of indistinguishability.

We have revisited topics such as information, uncertainty and their intrinsic relationships under a new point of view centered in the concept of indistinguishability.

In accordance with the road map described in the introductory chapter, emphasis has been put on providing applications for the theoretical contributions.

Let us describe more precisely the main contributions achieved as well as open issues and future lines of research.

Chapter 2.

Dempster-Shafer Theory of Evidence, as a framework for representing and managing general evidences, implicitly conveys the notion of indistinguishability between the elements of the domain of discourse based on their relative compatibility with the evidence at hand.

In chapter 2 we have mainly been concerned with providing definitions for the T-indistinguishability operator associated to a given body of evidence.

A first approach was defining operator E_1 as the T-indistinguishability operator generated by the one-point coverage function.

$$\forall x, y \in X : E_1(x, y) = \overleftrightarrow{T}(\mu_m(x), \mu_m(y)).$$

On the other hand, managing belief functions involve expensive computation both in terms of cost and storage. It thus makes sense to provide simpler approximations that are better suited to computation and explanation.

Basically, there are two main approaches to simplify a given belief function: reducing the number of focal elements, and constraining the evidence to belong to a predefined class having a relatively simple form. Two classes of evidence stand as obvious "simple" candidates: possibility and probability measures.

The following relation R over the set of *bpa* M

$$\forall m, m' \in M : (m, m') \in R \Longleftrightarrow \forall x, y \in X : E_1^m(x, y) = E_1^{m'}(x, y)$$

is an equivalence relation. We suggested a new approach to the problem of belief function approximation based on the partition on M induced by R.

For a given a basic probability assignment m we have proposed selecting another basic probability assignment from the same class of equivalence of m, having the singletons set as the set of focal elements (probabilistic approximation) or having nested focal sets (possibilistic approximation).

Unfortunately, as shown by example (2.3.1), uniqueness of the possibilistic approximation must be discarded.

Moreover, certain areas of application (like decision-making problems) require not just the relative concept of indistinguishability to be preserved by candidate approximations. The notion of ordering, which allows ranking the different alternatives under consideration, is then a fundamental issue so that substituting T-indistinguishability operators by T-preorders as the appropriate mathematical instrument seemed to be in order.

More formally, we rewrote relation R as

$$\forall m, m' \in M : (m, m') \in R_2 \Longleftrightarrow \forall x, y \in X : P_{\mu_m}(x, y) = P_{\mu_{m'}}(x, y)$$

where P_{μ_m} and $P_{\mu_{m'}}$ are the T-preorders generated by the one-point coverage function of m and m', respectively.

Theorem (2.3.19) proved that for any *bpa* m it exists a unique m' with nested focal sets, holding $(m, m') \in R_2$. In addition, it provides a constructive method to compute the possibilistic approximation m'.

We would like to point out that unlike other methods which assume the fulfillment of additional conditions (consistency, ...), our method does not impose any restriction over the original *bpa* m.

Despite of the fact that operator E_1 satisfies some intuitive requirements, an inherent drawback, as in any approximation-based approach, is the possible loss of information with respect to the original evidence. Hence, we introduced a general theorem providing in a constructive manner as well, the T-indistinguishability operator associated to any function: $F : \wp(X) \to [0, 1]$ and preserving as much of the information content as possible.

The generality of this theorem allows its specialization to a huge range of functions, particularly belief functions. In this case, the resulting T-indistinguishability operator (E_2) could be considered the natural T-indistinguishability operator providing the underlying indistinguishability relation to which the distribution of belief is committed.

$$\forall x, y \in X : E_2(x,y) = \min_{A \subset \wp(X-\{x,y\})} \overleftrightarrow{T}\big(Bel(\{x\} \cup A), Bel(\{y\} \cup A)\big).$$

In section 2.7 we have shown how theorem (2.4.2) could be applied to the field of Game Theory in order to compute the degree of indistinguishability among players in a cooperative game.

The characterization of one-dimensional E_2 operators has also been addressed, since this class of operators, in addition to affording greater clarity to the structure of the operator itself and significantly reducing the cost of computation, also enable their approximation by a single feature (generator) that carries exactly the same information from the point of view of indistinguishability as the original evidence.

Future work.

- Complete the study of relations R_1, R_2 and R_3 (see section 2.3.3) and providing necessary and sufficient conditions which a given pair of basic probability assignments must fulfill in order to belong to R_1, R_2 and R_3.
- Provide a full characterization of one-dimensional basic probability assignments, extending the achieved results in characterizing essential one-dimensionality.
- Extend the application of theorem (2.4.2) to the field of Game Theory and provide new applications to novel areas.

Chapter 3.

The study of methods for the measurement of uncertainty has grown in parallel with the acceptance that dealing with uncertainty has turned into a must for informational systems.

After providing a comprehensive summary of the state of the art on measures of uncertainty, we focused on tackling the problem of computing entropy when an indistinguishability relation has been defined over the elements of the domain. Then, entropy should be measured not according to the occurrence of different events, but according to the variability perceived by an observer equipped with indistinguishability abilities defined by the indistinguishability relation considered.

This interpretation introduced "the observer paradigm" which in turn lead to formalize the possibility of observing an element x_i (observation degree) as the sum of the probability that x_i really would have happened, plus the probability of having observed x_i mistakenly due to the occurrence of some other element x_j similar to x_i.

$$\pi(x_i) = p(x_i) + \sum_{x_j \in X | x_j \neq x_i} p(x_j) \cdot E(x_i, x_j).$$

We defined the observational entropy as the expected value of the observation degrees when they are measured in observable bits.

After proving interesting properties of observational entropy we defined the simultaneous observation degree, measuring the disagreement in the observations of an event by two different observers equipped with the same indistinguishability so that, for instance, observer A may claim having observed x_i while observer B has observed x_j, if x_i and x_j are "enough" indistinguishable.

A different situation arises when we consider observers with different indistinguishability abilities (each observer has its own indistinguishability operator). Then we defined the concept of conditional observation degree and conditional observation entropy based on the fact that observations performed by an observer A restrict the set of events that really might have occurred, affecting as well the variability of the potential observations of another observer B.

When both observers are equipped with indistinguishability operators that are equivalence relations then the conditional observational entropy particularizes to the classical heuristic function used by Quinlan in their classical algorithm for building decision trees.

Finally, joint observational entropy is also defined, and the classical law of total entropies is generalized to the observational paradigm as proved in theorem (3.6.55).

Future work.

- Extend the results and definitions related to the concept of observation entropy to continuous domains.
- Generalize the classical notion of independence to the observational paradigm.
- Provide an axiomatic characterization for observational entropy.

Chapter 4.

Real data is often pervaded with uncertainty so that devising techniques intended to induce knowledge in the presence of uncertainty seems entirely advisable.

The paradigm of computing with words follows this line in order to provide a computation formalism based on linguistic labels in contrast to traditional numerical-based methods. The use of linguistic labels enriches the understandability of the representation language, although it also requires adapting the classical inductive learning procedures to cope with such labels.

Among the existing methods, decision trees have become one of the most relevant paradigm within the machine learning community, mainly because of their proved applicability to a broad range of problems in addition to features as the readability of the knowledge induced. Variants of the original scheme proposed by Quinlan have been developed, providing decision trees with a more flexible methodology in order to cope with different kind of uncertainty.

Nevertheless, it is our opinion that these methods would benefit from the definition of an homogeneous framework since most of their particularities could be easily described as particularizations of more general procedures.

Chapter 4 has been devoted to describe our proposal for such a common framework. Furthermore, characterizations (in terms of the proposed framework) of relevant existing methods for inducing decision trees in the presence of uncertainty have been also provided.

A novel approach to building decision trees was introduced, addressing the case when uncertainty arises as a consequence of considering a more realistic setting in which decision maker's discernment abilities are taken into account when computing impurity measures. This novel paradigm resulted in what have been called "observational decision trees" since the main idea stems from the notion of observational entropy in order to incorporate indistinguishability concerns.

Among the existing methods for inducing rules from data, the sequential covering family of algorithms has a long tradition. In this chapter we have presented an algorithm (FSQ) intended to induce linguistic rules from data by properly managing the uncertainty present either in the set of describing labels or in the data itself. A formal comparison with other methods showed that the performance of FSQ is comparable with that of standard algorithms.

In the context of approximate reasoning or fuzzy control is usual to infer indistinguishability degrees between the elements of a given domain of discourse X, from indistinguishability operators defined on a set of fuzzy subsets of X. The opposite problem has not been paid the attention we think it deserves. In chapter 4 we have provided several techniques addressing this issue, related to the duality principle and the methods used in approximate reasoning.

Future work.

- Automatic "fitting" of labels (fuzzy templates) associated to the set of attributes depending on data distribution.
- Extend the generalized framework for induction of decision trees by dealing with fuzzy valued measures of compatibility.
- Post processing of syntactic simplification of the resulting set of rules induced by the algorithm FSQ.
- Application of techniques of pre-prunning and post-prunning in order to improve the accuracy and simplicity of observational decision trees.
- Enrich the language of representation of rules allowing the use of linguistic quantifiers.

Bibliography

[1] J. Abellan and S. Moral. A non-specificity measure for convex sets of probability distributions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8(3):357–367, 2000.

[2] J. Baldwin, J. Lawry, and P. Martin. Mass assignment based induction of decision trees on words. In *Proceedings of the VII Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, volume 1, pages 524–531, 1998.

[3] M. Bauer. Approximation algorithms and decision making in the Dempster-Shafer theory of evidence: an empirical study. *International Journal of Approximate Reasoning*, (17):217–237, 1996.

[4] M. Berthold. Mixed fuzzy rule formation. *International Journal of Approximate Reasoning*, (32):67–84, 2003.

[5] J.C Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, 1981.

[6] J.C. Bezdek and J.O. Harris. Fuzzy partitions and relations: an axiomatic basis for clustering. *Fuzzy Sets and Systems*, (1):112–127, 1978.

[7] M. Bjanger. Induction of decision trees from partially classied data using belief functions. Master's thesis, Norwegian University of Science and Technology, 2000.

[8] S. Bodjanova. Approximation of fuzzy concepts in decision making. *Fuzzy Sets and Systems*, (85):23–29, 1997.

[9] D. Boixader. *Contribució a l'estudi dels morfismes entre operadors d'indistingibilitat. Aplicació al raonament aproximat*. PhD thesis, Universitat Politècnica de Catalunya, 1997.

[10] D. Boixader and J. Jacas. Generators and dual T-indistinguishabilities. In B. Bouchon-Meunier, R. Yager, and

L. Zadeh, editors, *Fuzzy Logic and Soft Computing*, pages 283–291. World Scientific, 1994.

[11] D. Boixader, J. Jacas, and J. Recasens. Fuzzy equivalence relations: advanced material. In D. Dubois and H. Prade, editors, *Fundamentals of Fuzzy Sets*, pages 261–290. Kluwer, 1999.

[12] D. Boixader, J. Jacas, and J. Recasens. Upper and lower approximation of fuzzy sets. *International Journal of General Systems*, (29):555–568, 2000.

[13] B. Bouchon-Meunier and R. Yager. Entropy of similarity relations in questionnaires and decision trees. In *Proceedings of the II International Conference on Fuzzy Systems*, pages 1225–1230, 1993.

[14] Bratko, Cestnik, and Kononenko. *Assistant 86: A knowledge-elicitation tool for sophisticated users.* Sigma Press, 1986.

[15] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[16] L. Breiman et al. *Classification and regression trees.* Wadsworth International Group, 1984.

[17] T. Calvo, G. Mayor, and R. Mesiar. *Aggregation operators: new trends and applications.* Physica-Verlag GmbH, 2002.

[18] J. Casillas, O. Cordon, and F. Herrera. Learning fuzzy rules using ant colony optimization algorithms. In *Proceedings of the II International Workshop on Ant Algorithms (ANTS)*, pages 13–21, 2000.

[19] J. Cendrowska. Prism: an algorithm for inducing modular rules. *International Journal of Man-Machine Studies*, (27):349–370, 1987.

[20] K. Chan and W. Au. An effective algorithm for discovering fuzzy rules in relational databases. Number vol2, pages 1314–1319, 1998.

[21] V. Cherkassky and M. Mulier. *Learning from data.* John Wiley and Sons, 1998.

[22] P. Clark and R. Niblett. The CN2 induction algorithm. *Machine Learning*, (3):261–284, 1989.

[23] W. Cohen. Fast effective rule induction. In A. Prieditis and S. Russell, editors, *Proceedings of the XII International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann, 1995.

[24] A. De Luca and S. Termini. A definition of a non probabilistic entropy in the setting of fuzzy sets theory. *Information and Control*, (20):301–312, 1972.

[25] A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, (38):325–339, 1967.

[26] T. Denoeux. Inner and outer clustering approximations of belief functions. In *Proceedings of the VIII Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, pages 125–132, 2000.

[27] T. Denoeux and A. Ben Yaghlane. Approximating the combination of belief functions using the fast Möbius transform in a coarsened frame. *International Journal of Approximate Reasoning*, (31):77–101, 2002.

[28] M. Drobics, U. Bodenhofer, and P. Klement. FS-FOIL: an inductive learning method for extracting interpretable fuzzy descriptions. *International Journal of Approximate Reasoning*, (32):131–152, 2003.

[29] D. Dubois and H. Prade. On several representations of an uncertain body of evidence. In Gupta M. and Sanchez E., editors, *Fuzzy Information and Decision Processes*, pages 279–283. North Holland, 1982.

[30] D. Dubois and H. Prade. Fuzzy cardinality and the modelling of imprecise quantification. *Fuzzy Sets and Systems*, (16):199–230, 1985.

[31] D. Dubois and H. Prade. A note on measures of specifity for fuzzy sets. *International Journal of General Systems*, (10):279–283, 1985.

[32] D. Dubois and H. Prade. Properties of measures of information in evidence and possibility theories. *Fuzzy Sets and Systems*, (24):161–182, 1987.

[33] D. Dubois and H. Prade. Consonant approximations of belief functions. *International Journal of Approximate Reasoning*, (4):419–449, 1990.

[34] D. Dubois and H. Prade. Rough fuzzy sets and fuzzy rough sets. *International Journal of General Systems*, (17):191–209, 1990.

[35] D. Dubois and H. Prade. Fuzzy sets and probability: misunderstandings, bridges and gaps. In *Proceedings of the II International Conference on Fuzzy Systems*, pages 1059–1068, 1993.

[36] D. Dubois, H. Prade, and E. Rannou. User driven summarization of data based on gradual rules. In *Proceedings of the VI International Conference on Fuzzy Systems*, number Vol2, pages 839–844, 1997.

[37] Z. Elouedi, K. Mellouli, and P. Smets. Decision trees using the belief function theory. In *Proceedings of the VIII Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, 2000.

[38] F. Esposito, M. Malerba, and G. Someraro. A comparative analysis of methods for prunning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):476–491, 1997.

[39] Fazlollah and M. Reza. *An introduction to Information Theory.* Dover Publications, 1994.

[40] E. Frank and I. Witten. Generating accurate rule sets without global optimization. In *Proceedings of the XV International Conference of Machine Learning*, 1998.

[41] L. Garmendia. *Contribución al estudio de las medidas en la lógica borrosa: condicionalidad, especifidad y transitividad.* PhD thesis, Universidad Politécnica de Madrid, 2001.

[42] L. Garmendia, R. Yager, E. Trillas, and A. Salvador. On t-norms based measures of specificity. *Fuzzy Sets and Systems*, 133(2):237–248, 2003.

[43] J. Gebhardt, F. Klawoon, and R. Kruse. *Foundations of Fuzzy Systems.* John Wiley.

[44] P. Gil. *Teoría Matemática de la Información.* Ediciones ICE, 1981.

[45] Ll. Godo. *Contribució a l'estudi de models d'inferència en els sistemes possibilístics.* PhD thesis, Universitat Politècnica de Catalunya, 1990.

[46] J.R. Goodman. Fuzzy sets as equivalence classes of random sets. In Yager R., editor, *Fuzzy Sets and Possibility Theory*, pages 327–342. Pergamon Press, 1982.

[47] J. Gordon and E. Shortliffe. A method of managing evidential reasoning in a hierarchical hypothesis space. *Artificial Intelligence*, (26):323–357, 1985.

[48] M. Grabisch and M. Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal on Game Theory*, 4(28):547–565, 1999.

[49] R. Gregson. *Psychometrics of Similarity.* Academic Press, 1988.

[50] R. Haenni and N. Lehmann. Resource bounded and anytime approximation of belief fuunction computations. *International Journal of Approximate Reasoning*, (31):103–154, 2002.

[51] D. Harmanec. Faithful approximations of belief functions. In *Proceedings of the XV Conference of Uncertainty in Artificial Intelligence*, pages 13–21, 1999.

[52] R. Hartley. Transmission of information. *The Bell System Technical Journal*, (7):535–563, 1928.

[53] J. Herencia and M. Lamata. Entropy measures associated with a fuzzy basic probability assignment. In *Proceedings of the VI International Conference on Fuzzy Systems*, pages 863–868, 1997.

[54] E. Hernández and J. Recasens. Un model lingüístic per la descripció de conceptes. In *Actes del I Congrés Català d'Intel.ligència Artificial (CCIA)*, pages 181–184, 1998.

[55] E. Hernández and J. Recasens. Un modelo lingüístico para la descripción de conceptos. In *Actas del VIII Congreso sobre tecnologías y lógica fuzzy (ESTYLF)*, pages 103–107, 1998.

[56] E. Hernández and J. Recasens. Observational entropy: entropy in the context of indistinguishability operators. In *Proceedings of the I Conference of the European Society of Fuzzy Logic and Technology (EUSFLAT)*, pages 203–266, 1999.

[57] E. Hernández and J. Recasens. Entropy and indistinguishability: observational entropy. In *Proceedings of the VIII Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, pages 870–875, 2000.

[58] E. Hernández and J. Recasens. A fuzzy sequential covering algorithm for the generation of rules. In *Actas del X Congreso sobre tecnologías y lógica fuzzy (ESTYLF)*, pages 319–322, 2000.

[59] E. Hernández and J. Recasens. Growing decision trees in presence of indistinguishability: observational decision trees. In *Proceedings of the II Conference of the European Society of Fuzzy Logic and Technology (EUSFLAT)*, 2001.

[60] E. Hernández and J. Recasens. Generating indistinguishability operators from prototypes. *International Journal of General Systems*, 17(12):1131–1142, 2002.

[61] E. Hernández and J. Recasens. On possibilistic and probabilistic approximations of unrestricted belief functions based on the concept

of fuzzy T-preorder. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(2):185–200, 2002.

[62] E. Hernández and J. Recasens. A reformulation of entropy in the presence of indistinguishability operators. *Fuzzy Sets and Systems*, 128(2):185–196, 2002.

[63] E. Hernández and J. Recasens. Relaciones de indistinguibilidad en el marco de la teoría de la evidencia. In *Actas del XI Congreso sobre tecnologías y lógica fuzzy (ESTYLF)*, pages 367–371, 2002.

[64] E. Hernández and J. Recasens. A general framework for induction of decision trees under uncertainty. In J. Lawry, J. Shanahan, and A. Ralescu, editors, *Modelling with Words. LNAI 2873*, pages 26–43. Springer, 2003.

[65] E. Hernández and J. Recasens. Indistinguishability relations in Dempster-Shafer theory of evidence. *International Journal of Approximate Reasoning*, 37(3):145–187, 2004.

[66] E. Hernández and J. Recasens. Indistinguishability in cooperative games. In *Proceedings of the IV Conference of the European Society of Fuzzy Logic and Technology (EUSFLAT)*, 2005.

[67] E. Hernández and J. Recasens. Modelling T-indistinguishability in game theory. In *Proceedings of the XI Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, 2006.

[68] F. Herrera and M. Lozano. Fuzzy genetic algorithms: issues and models.

[69] M. Higashi and G. Klir. Measures of uncertainty and information based on possibility distributions. *International Journal of General Systems*, (9):43–58, 1983.

[70] K. Hirota and W. Pedrycz. Linguistic data mining and fuzzy modelling. *Proceedings of the V International Conference on Fuzzy Systems, year=1996, pages=1488-1492.*

[71] U. Höhle. Entropy with respect to plausibility measures. In *Proceedings of the XII IEEE International Symposium of Multiple-Valued Logic*, pages 167–169, 1982.

[72] U. Höhle. Quotients with respect to similarity relations. *Fuzzy Sets and Systems*, (27):31–44, 1988.

[73] R. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1):63–90, 1993.

[74] Hong, Wang, and Tseng. Inductive learning from fuzzy examples. *Proceedings of the V International Conference on Fuzzy Systems*, (vol1):13–18, 1996.

[75] H. Ishibuchi, K. Nozaki, H Tanaka, Y Hosaka, and M Matsuda. Empirical study on learning in fuzzy systems by rice test analysis. *Fuzzy Sets and Systems*, (64):129–144, 1994.

[76] J. Jacas. *Contribució a l'estudi dels operadors d'indistingibilitat i a les seves aplicacions als processos de classificació*. PhD thesis, Universitat Politècnica de Catalunya, 1987.

[77] J. Jacas. On the generators of a T-indistinguishability operator. *Stochastica*, pages 49–63, 1988.

[78] J. Jacas and J. Recasens. Fuzzy numbers and equality relations. In *Proceedings of the II International Conference on Fuzzy Systems*, pages 1298–1301, 1993.

[79] J. Jacas and J. Recasens. Fixed points and generators of fuzzy relations. *Journal of Mathematical Analysis and Applications*, (186):21–29, 1994.

[80] J. Jacas and J. Recasens. Fuzzy T-transitive relations: eigenvectors and generators. *Fuzzy Sets and Systems*, (72):147–154, 1995.

[81] J. Jacas and J. Recasens. Maps between indistinguishablity operators. In *Proceedings Joint IX IFSA World Congress and XX NAFIPS International Conference*, pages 1171–1175, 2001.

[82] C. Janikow. Fuzzy decision trees: issues and methods. *IEEE Transactions on Systems, Man and Cybernetics*, 28(1):1–14, 1998.

[83] C. Joslyn and G. Klir. Minimal information loss possibilistic approximations of random sets. In *Proceedings of the I International Conference on Fuzzy Systems*, pages 1081–1088, 1992.

[84] J. Kacprzyk and C. Iwanski. Fuzzy logic with linguistic quantifiers in inductive learning. In L. Zadeh and J. Kacprzyk, editors, *Fuzzy logic for the management of uncertainty*, pages 465–478. Wiley, 1992.

[85] M. Kamber and R. Shinghal. Evaluating the interestingness of characteristic rules. In *Proceedings of the II International Conference on Knowledge Discovery and Data Mining*, pages 263–266, 1996.

[86] J. Kampe de Feriet. Interpretations of membership functions of fuzzy sets in term of plausibility and belief. In Gupta M. and Sanchez E., editors, *Fuzzy Information and Decisions Processes*, pages 93–98. North Holland, 1982.

[87] J. Kaufmann. Introduction to the theory of fuzzy sets. In *New York Academic Press*. 1975.

[88] R. Kennes. Computational aspects of the Möbius transform of graphs. *IEEE Transactions on Systems, Man and Cybernetics*, (22):201–223, 1992.

[89] R. Kennes and P. Smets. Computational aspects of the Möbius transformation. In *Uncertainty in Artificial Intelligence*, pages 401–416, 1991.

[90] F. Klawonn and R. Kruse. Derivation of fuzzy classification rules from multidimensional data. In Lasker and Liu, editors, *Advances in Intelligent Data Analysis*, pages 90–94. Windsor, 1995.

[91] F. Klawoon and R. Kruse. Equality relations as a basis for fuzzy control. *Fuzzy Sets and Systems*, (54):147–156, 1993.

[92] G. Klir. Where do we stand on measures of uncertainty, ambiguity, fuzziness and the like? *Fuzzy Sets and Systems*, (24):141–160, 1987.

[93] G. Klir. Uncertainty and information measures for imprecise probabilities: an overview. In *Proceedings of the I International Symposium on Imprecise Probabilities and Their Applications*, 1999.

[94] G. Klir and A. Ramer. Uncertainty in Dempster-Shafer theory: a critical re-examination. *International Journal of General Systems*, (18):155–166, 1990.

[95] G. Klir and R. Smith. On measuring uncertainty and uncertainty-based information: recent developments. *Annals of Mathematics and Artificial Intelligence*, (32):5–33, 2001.

[96] G Klir and M. Wierman. *Uncertainty based information. Elements of generalized information theory*. Physica-Verlag, 1999.

[97] G. Klir and B. Yuan. *Fuzzy sets and fuzzy logic: theory and applications*. Prentice Hall, 1995.

[98] R. Kohavi. The power of decision tables. In Nada Lavrac and Stefan Wrobel, editors, *Proceedings of the European Conference on Machine Learning*, Lecture Notes in Artificial Intelligence 914, pages 174–189. Springer Verlag, 1995.

[99] B. Kosko. Fuzzy entropy and conditioning. *Information Science*, (40):165–174, 1986.

[100] R. Kruse and D. Nauck. Learning methods for fuzzy systems. In *Proceedings of Fuzzy–Neuro–Systems*, pages 7–22, 1995.

[101] M. Lamata and S. Moral. Measures of entropy in the theory of evidence. *International Journal of General Systems*, (14):297–305, 1987.

[102] D. Lee and Kim M. Database summarization using fuzzy ISA hierarchies. *IEEE Transactions on Systems Man and Cybernetics*, 27(4):671–680, 1997.

[103] C.H. Ling. Representation of associative functions. *Math. Debrecen*, (12):182–212, 1965.

[104] Liu, Wang, Hong, and Tseng. A fuzzy inductive learning strategy for modular rules. *Fuzzy Sets and Systems*, (103):91–105, 1999.

[105] D. Loutchmia and H. Ralambondrainy. Inductive learning using similarity measures on lattice fuzzy sets. In *Proceedings of the VI International Conference on Fuzzy Systems*, pages 1307–1313, 1997.

[106] J. Lowrance, T. Garvey, and T. Strat. A framework for evidential reasoning systems. In *Proceedings of the V National Conference of the American Association for Artificial Intelligence*, pages 896–903, 1986.

[107] P.E Maher and D. Saint-Clair. Uncertain reasoning in an ID3 machine learning framework. In *Proceedings of the II International Conference on Fuzzy Systems*, pages 7–12, 1993.

[108] R. Mantaras. A distance-based attribute selection measure for decision tree induction. *Machine learning*, 6(1):81–92, 1991.

[109] J.L. Marichel. *Aggregation operators for multicriteria decision aid.* PhD thesis, Universite de Liege, 1998.

[110] C. Marsala. *Inductive learning in presence of imprecise data: Methods to build and to use fuzzy decision trees.* PhD thesis, Pierre et Marie Curie University, 1998.

[111] B. Martin. *Instance based learning: nearest neighbor with generalization.* PhD thesis, University of Waikato, 1995.

[112] K. Menger. Probabilistic theories of relations. In *Proceedings of the National Academy of Sciences*, volume 37, pages 178–180, 1951.

[113] M. Michalski. On the quasi-minimal solution of the general covering problem. In *Proceedings of the I International Symposium on Information Processing*, pages 125–128, 1969.

[114] R. Michalski. A theory and methodology of inductive learning. *Artificial Intelligence*, (20):111–161.

[115] J. Mingers. An empirical comparison of prunning methods for decision tree induction. *Machine Learning*, (4):227–243, 1989.

[116] J. Mingers. An empirical comparison of selection measures for decision tree induction. *Machine Learning*, (3):319–342, 1989.

[117] T. Mitchell. *Version spaces: an approach to concept learning.* PhD thesis, Stanford University, 1979.

[118] T. Mitchell. Generalization as search. *Artificial Intelligence*, (18):203–224, 1982.

[119] T. Mitchell. *Machine Learning.* Mc. Graw-Hill, 1997.

[120] S. Murthy. *On growing better decision trees from data.* PhD thesis, John Hopkins University, 1995.

[121] D. Newman, S. Hettich, C. Blake, and C. Merz. UCI repository of machine learning databases, 1998.

[122] H. Nguyen. *On entropy of random sets and possibility distributions.* CRC Press, 1985.

[123] C. Olaru and L. Wehenkel. A complete fuzzy decision tree technique. *Fuzzy Sets and Systems*, (138):221–254, 2003.

[124] S.V. Ovchinnikov. The duality principle in fuzzy set theory. *Fuzzy Sets and Systems*, (42):133–144, 1991.

[125] R. Pal. On quantification of different facets of uncertainty. *Fuzzy Sets and Systems*, (107):81–91, 1999.

[126] R. Pal and J. Bezdek. Measuring fuzzy uncertainty. *IEEE Transactions on Fuzzy Systems*, 2(2):107–118, 1994.

[127] R. Pal, J. Bezdek, and R. Hemasinha. Uncertainty measures for evidential reasoning. part I: a review. *International Journal of Approximate Reasoning*, (7):165–183, 1992.

[128] R. Pal, J. Bezdek, and R. Hemasinha. Uncertainty measures for evidential reasoning. part II: a new measure of total uncertainty. *International Journal of Approximate Reasoning*, (8):1–16, 1993.

[129] Z. Pawlak. *Rough Sets.* Kluwer Academic Publishers, 1991.

[130] W. Pedrycz. Data mining and knowledge discovery: a fuzzy set perspective. *Tatra Mountains Math. Publ*, (13):195–218, 1997.

[131] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetstky-Shapiro and W. Frawley, editors, *Knowledge discovery in databases*. MIT Press, 1991.

[132] H. Poincaré. *La science et l'hypothése*. Flammarion, 1902.

[133] J.R Quinlan. Induction of decision trees. *Machine Learning*, pages 81–106, 1986.

[134] J.R Quinlan. Probabilistic decision trees. In *Machine Learning*, pages 140–152, 1990.

[135] J.R Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann, 1993.

[136] J. Recasens. *Sobre la representació i generació de relacions dindistingibilitat*. PhD thesis, Universitat Politècnica de Catalunya, 1992.

[137] J. Recasens and A. de Soto. Modelling a linguistic variable as a hierarchical family of partitions induced by an indistinguishability operator. *Fuzzy Sets and Systems*, 121(3):427–437, 2001.

[138] F. Reza. *An Introduction to Information Theory*. Dover Publications.

[139] A. Ribas et al. *Aprendizaje automático*. Edicions UPC, 1994.

[140] J. Rives. FID3: fuzzy decision tree. In *Proceedings of the I International Symposium of Uncertainty, Modelling and Analysis*, pages 457–462, 1990.

[141] E. Ruspini. A theory of fuzzy clustering. In *Proceedings of the IEEE Conference on Decision and Control*, pages 1378–1383, 1977.

[142] E. Ruspini. Recent developments in fuzzy clustering. In Yager R., editor, *Fuzzy set and possibility theory: recent developments*, pages 133–147. Pergamon Press, 1982.

[143] T. Sales. Fuzzy sets as set classes. *Stochastica*, (6):249–264, 1982.

[144] M. Sarkar. Rough-fuzzy functions in classification. *Fuzzy Sets and Systems*, (132):353–369, 2002.

[145] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[146] G. Shafer and R. Logan. Implementing Dempster's rule for hierarchical evidence. *Artificial Intelligence*, (33):271–298, 1987.

[147] C. Shannon and W. Weber. *The mathematical theory of communication*. University of Illinois Press, 1964.

[148] L. Shapley. A value for n-person games. In H. Kuhn and A. Tucker, editors, *Contributions to Game Theory*, pages 307–317. Princeton University Press, 1953.

[149] P. Smets. Information content of an evidence. *International Journal of Man-Machine Studies*, (19):33–43, 1983.

[150] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, (66):191–234, 1994.

[151] R. Smith. *Generalized information theory: resolving some old questions and opening some new ones*. PhD thesis, Binghampton University, 2000.

[152] M. Sugeno. Fuzzy measures and fuzzy integrals: a survey. In *Fuzzy automata and decision process*, pages 89–102. North Holland, 1977.

[153] B. Tessem. Approximations for efficient computation in the theory of evidence. *Artificial Intelligence*, (61):315–329, 1993.

[154] E. Trillas. Assaig sobre les relacions d'indistingibilitat. In *Actes del I Congrés Català de Lògica Matemàtica*, pages 51–59, 1982.

[155] E. Trillas and C. Alsina. Introducción a los espacios métricos generalizados. In *Serie Universitaria*. Fund. Juan March, 1979.

[156] M. Umano et al. Fuzzy decision trees by using fuzzy ID3 algorithm and its application to diagnosis systems. In *Proceedings of the III International Conference on Fuzzy Systems*, pages 2113–2118, 1994.

[157] P.E. Utgoff and J.A. Clouse. A Kolmogorov-Smirnoff metric for decision tree induction. Technical Report 96-3, University of Massachusets, 1996.

[158] L. Valverde. On the structure of F-indistinguishability operators. *Fuzzy Sets and Systems*, (17):313–328, 1985.

[159] Van de Merckt. Decision trees in numerical attribute spaces. In *Proceedings of the XIII International Joint Conference on Artificial Intelligence*, pages 1016–1021, 1993.

[160] J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[161] F. Voorbraak. A computationally efficient approximation of Dempster-Shafer theory. *International Journal of Man-Machine studies*, (30):525–536, 1989.

[162] L. Wang and J. Mendel. Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man and Cybernetics*, 22(6):1414–1427, 1992.

[163] R. Weber. Fuzzy-ID3: a class of methods for automatic knowledge acquisition. In *Proceedings of the II International Conference on Fuzzy Logic and Neural Networks*, pages 265–268, 1992.

[164] N. Wilson. Algorithms for Dempster-Shafer theory. In D. Gabbay and P. Smets, editors, *Handbook of defeasible reasoning and uncertainty management*, pages 421–475. Kluwer, 2000.

[165] S. Wilson and S. Moral. Fast Markov chain algorithms for calculating Dempster-Shafer belief. In *Proceedings of the XII European Conference on Artificial Intelligence*, pages 672–678, 1996.

[166] I. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations.* Morgan Kaufmann.

[167] W. Xie and D. Bedrosian. An information measure for fuzzy sets. *IIIE Transactions on Systems, Man and Cybernetics.*, 14:151–156, 1984.

[168] R. Yager. On the specifity of a posibility distribution. *Fuzzy Sets and Systems*, 50(3).

[169] R. Yager. On the measure of fuzziness and negation. part I: membership in the unit interval. *International Journal of General Systems*, (5):221–229, 1979.

[170] R. Yager. Measuring tranquility and anxiety in decision making: an application of fuzzy sets. *International Journal of General Systems*, (8):139–146, 1982.

[171] R. Yager. Entropy and specificity in a mathematical theory of evidence. *International Journal of General Systems*, (9):249–260, 1983.

[172] R. Yager. Toward a general theory of reasoning under uncertainty. part I: non-specificity and fuzziness. *International Journal of General Systems*, (1):45–67, 1986.

[173] R. Yager. Toward a general theory of reasoning under uncertainty. part II: probability. *International Journal of Man-Machine Studies*, (25):613–631, 1986.

[174] R. Yager. On linguistic summaries of data. In G. Piatetstky-Shapiro and W. Frawley, editors, *Knowledge discovery in databases*, pages 347–363. MIT Press, 1991.

[175] R. Yager. Similarity based specificity measures. *International Journal of General Systems*, 19:91–105, 1991.

[176] R. Yager. Default knowledge and measures of specificity. *Information Science*, (61):1–44, 1992.

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

[178] R. Yager. Counting the number of classes in a fuzzy set. *IEEE Transactions on Systems, Man and Cybernetics*, 23(1):257–264, 1993.

[179] R. Yager. Fuzzy summaries in database mining. In *Proceedings of the IV International Conference on Fuzzy Systems*, pages 265–269, 1995.

[180] R. Yager. A class of fuzzy measures generated from a Dempster-Shafer belief structure. *International Journal of General Systems*, (14):1239–1247, 1999.

[181] R. Yager. On the entropy of fuzzy measures. *IEEE Transactions on Fuzzy Systems*, 8(4):453–461, 2000.

[182] R. Yager and D. Rasmussen. Summary-SQL: a fuzzy tool for data mining. *Intelligent Data Analysis*, 1(1):49–58, 1997.

[183] R. Yager and Rubinson T. Linguistic summaries of databases. In *Proceedings of IEEE Conference on Decision and Control*, pages 1094–1097, 1981.

[184] J. Yen. Generalizing the Dempster-Shafer theory to fuzzy sets. *IEEE Transactions on Systems, Man and Cybernetics*, 20(3):559–569, 1990.

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

[186] L. Zadeh. Fuzzy sets. *Inf. Control*, (8):338–353, 1965.

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

[188] L. Zadeh. Similarity relations and fuzzy orderings. *Information Sciences*, (3):177–200, 1971.

[189] L. Zadeh. The concept of a linguistic variable and its applications to approximate reasoning I, II and III. In *Fuzzy Sets and Applications: selected papers by L.A.Zadeh*, pages 218–366, 1976.

[190] L. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1):3–28, 1978.

[191] L. Zadeh. Fuzzy sets and information granularity. In M. Gupta, R. Ragade, and R. Yager, editors, *Advances in Fuzzy Set theory and applications*, pages 3–18. North Holland, 1979.

[192] L. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computing and Mathematics with Applications*, (9):149–184, 1983.

[193] L. Zadeh. Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems*, 4(2):103–111, 1996.

[194] L. Zadeh. Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. *Journal of Statistical Planning and Inference*, (105):233–264, 2002.