Representations of Uncertainty in AI: Beyond Probability and Possibility

Thierry Denoeux, Didier Dubois, Henri Prade

Representations of Uncertainty in AI: Beyond Probability and Possibility

Thierry Denœux, Didier Dubois, and Henri Prade

Abstract This chapter completes the survey of the existing frameworks for representing uncertain and incomplete information, started in the previous chapter of this volume. The theory of belief functions and the theory of imprecise probabilities are presented. The latter setting is mathematically more general than the former, and both include probability theory and quantitative possibility theory as particular cases. Their respective knowledge representation capabilities are highlighted.

1 Introduction

Usually items of information are neither precise nor always coherent with one another. This chapter presents two uncertainty theories that generalize probability theory while being capable of handing incomplete information in an explicit way, by including possibility theory as a special case. There are two ways of building such a generalized framework.

The first idea is to introduce probability theory on top of the basic set-valued representation of incomplete information. Dempster imagined a set equipped with a probability distribution and a one-to-many mapping from this set to a space of interest. Such probabilities can be subjective or frequentist. Upper and lower probabilities are then obtained on the second space. Dempster considered this set-up as

Thierry Denoeux
Sorbonne Universités, Université de Technologie de Compiègne, CNRS, Heudiasyc (UMR 7253)
e-mail: tdenoeux@utc.fr

Didier Dubois
IRIT-CNRS, Université Paul Sabatier, Toulouse, France,
e-mail: dubois@irit.fr

Henri Prade
IRIT-CNRS, Université Paul Sabatier, Toulouse, France,
e-mail: prade@irit.fr

an extension of the fiducial paradigm for statistical inference, while Shafer interpreted these upper and lower probabilities as plausibility and belief functions without reference to an underlying probability space with a one-to-many mapping. The approach so-obtained was called theory of evidence by Shafer. It is tailored for the representation and merging of unreliable pieces of evidence. In contrast, upper and lower probabilities in Dempster set-up may also model ill-known probabilities due to incomplete observations of random variables.

The second idea is to work with (convex) sets of probabilities, either because the statistical model is ill-known, or because the usual protocol for generating subjective probabilities is altered, admitting that buying and selling prices of lotteries attached to risky events may differ. The latter is the basis of Walley theory of lower previsions and imprecise probabilities. It turns out that the framework of Walley is mathematically more general than the theory of Dempster-Shafer. This chapter provides an account of these generalizations of Bayesian probability theory.

2 Theory of Belief Functions

The belief function model [Shafer, 1976, 1990; Yager and Liu, 2008] adds probabilities on top of the set-based approach to imprecision. It replaces a representation of the form $v \in A$, where A is a set of possible values of v, by a discrete probability distribution over possible statements of the form $v \in A$ (assuming the universe, called *frame of discernment* by Shafer, S is finite). We denote by m such a probability distribution on the power set 2^S of S (the set of all subsets of S). As m is a probability distribution, condition $\sum_{A \subseteq S} m(A) = 1$ is verified. Function m is called a *mass function*, and $m(A)$ is called the *belief mass* assigned to subset A. Any subset A of S such that $m(A) > 0$ is called a *focal set* of m. We denote by $\mathscr{F}$ the family of focal sets. In general, we do not assign any positive mass to the empty set, i.e., we assume that $m(\emptyset) = 0$; mass function m is then said to be *normalized*. However, the Transferable Belief Model (TBM) [Smets and Kennes, 1994] relaxes this constraint: the mass $m(\emptyset)$ then represents the degree of internal contradiction of the mass function.

In this hybrid representation of uncertainty, it is important to understand the meaning of the mass function. In particular, the belief mass $m(A)$ should not be confused with a probability of occurrence of A. According to Shafer [1976], $m(A)$ is "the measure of the belief committed exactly to A". More precisely, we can say that $m(A)$ is the probability that the agent *only knows* that $v \in A$. There is thus an implicit epistemic modality in $m(A)$, which is absent from $P(A)$. This is the reason why function m may be non-monotonic with respect to inclusion: we may have $m(A) > m(B) > 0$ when $A \subset B$, if the agent is sure enough that what they know is of the form $v \in A$. In particular, $m(S)$ is the probability that the agent does not know anything. The *vacuous* mass function $m^?$ defined by $m^?(S) = 1$ thus represents total ignorance. This epistemic interpretation of mass functions is in line with Shafer [1981]'s *random code* metaphor outlined in the next section.

2.1 Random Code Semantics

A mass function can be interpreted by considering that the information provided by a source (a piece of evidence) can be assimilated to a coded message whose meaning is random [Shafer, 1981]. More precisely, assume that the source sends an encrypted message using a code chosen at random from a set $C = \{c_1, \ldots, c_n\}$ with probabilities $p_1, \ldots, p_n$. We know the set of codes as well as the chances of each code to be selected. If we decode the message using code c_i, we get a decoded message of the form $v \in \Gamma(c_i) = A_i$, where Γ is a multivalued mapping from C to 2^S. The probability that the meaning of the original message is $v \in A$ is thus

$$m(A) = \sum_{\{1 \leq i \leq n : A_i = A\}} p_i. \tag{1}$$

In particular, the probability that the message is empty, i.e., that it contains no information about v, is $m(S)$. The triple (C, P, Γ), where P is a probability measure on C, defines a random set [Nguyen, 2006]. The formal equivalence between random sets and belief functions has been proved for the first time by Nguyen [1978]. However, in random set theory, sets A with $m(A) > 0$ do not necessarily represent states of knowledge. They can be objects taking the form of sets [Couso et al, 2014], contrary to the case of evidence theory illustrated in the following example.

> **Example**: Consider a watch that may be out of order with some known probability ε. The set C describes the set of states of the watch, $C = \{\mathsf{working}, \mathsf{broken}\}$. Assume that the watch shows time h. In that case, the multivalued mapping Γ is $\Gamma(\mathsf{working}) = \{h\}$ (if the watch is working, it shows the right time), and $\Gamma(\mathsf{broken}) = S$ (if it is out of order, we do not know what time it is). The mass function induced by S is thus $m(\{h\}) = 1 - \varepsilon$ and $m(S) = \varepsilon$.

The mass function obtained in the previous example is said to be *simple* because the belief mass is shared between a single subset A of S, and S itself. Such a mass function arises when a non-reliable source states that $v \in A$, and the agent believes that the source is irrelevant with probability ε. This probability is committed to S whereas $m(A) = 1 - \varepsilon$.

This way of generating a mass function from a multivalued mapping was first proposed by Dempster [1967] in the context of statistical inference. Shafer [1976] renamed the upper and lower probabilities of Dempster *plausibility and belief functions*, respectively. To quote Shafer [2016b]'s recent intellectual autobiography:

> My thought was to surrender the word *probability* to the objective concept and to build a new subjective theory using mainly the word *belief*.

A mass function m models a state of knowledge, whereas the underlying triple (C, P, Γ) represents a piece of evidence with uncertain meaning. Among theories of uncertainty, the theory of belief functions has the particularity of putting emphasis on the evidence that generates a state of knowledge, as shown by the title of Shafer [1976]'s seminal book: *A Mathematical Theory of Evidence*.

2.2 Basic Set Functions

A mass function m induces two set functions: a belief function Bel (for "*belief*") and a plausibility function Pl, defined, respectively, by

$$Bel(A) = \sum_{E \subseteq A, E \neq \emptyset} m(E); \quad Pl(A) = \sum_{E \cap A \neq \emptyset} m(E). \tag{2}$$

Observe that $\forall A, Bel(A) \leq Pl(A)$. When $m(\emptyset) = 0$, it is clear that $Bel(S) = Pl(S) = 1, Pl(\emptyset) = Bel(\emptyset) = 0$, and $Bel(A) = 1 - Pl(\overline{A})$. Consequently, these two functions are dual, as are necessity and possibility functions. The degree of belief $Bel(A)$ can be interpreted as the probability of provability of A from the available knowledge represented by m. In the language of modal logic, we should write $Bel(A) = P(\Box A)$, where $\Box$ represents the modality of provability [Pearl, 1990]. In the same way, $Pl(A)$ can be seen as the probability of logical consistency of A with m.

Belief functions Bel are *completely monotone*, i.e., for any $k \geq 2$ and any family $(A_1, \ldots, A_k)$ of subsets of S, the following inequality holds,

$$Bel\left(\bigcup_{i=1,\ldots,k} A_i\right) \geq \sum_{i=1}^{k} (-1)^{i+1} \sum_{I:|I|=i} Bel\left(\bigcap_{j \in I} A_j\right). \tag{3}$$

For Shafer [2016b], these inequalities play for belief functions the same role as Kolmogorov axioms for probability theory. Plausibility functions verify a similar property (they are *completely alternating*), changing the direction of the inequality and switching the $\cap$ and $\cup$ operations.

A *commonality function*

$$Q(A) = \sum_{E \supseteq A} m(E) \tag{4}$$

was also introduced in [Shafer, 1976], essentially for computational reasons. It later appeared that the commonality function is an extension of the guaranteed possibility function in possibility theory [Dubois et al, 2001] (see the previous Chapter 3 in this volume).

Conversely, knowing function Bel, we can uniquely recover function m by the Möbius transform

$$m(E) = \sum_{A \subseteq E} (-1)^{|E \setminus A|} Bel(A).$$

Similar identities make it possible to recover m from Pl or Q. The fast Möbius transform [Kennes, 1992] can perform these operations efficiently.

Belief functions are often defined on finite universes. Yet, thanks to the formal identity between belief functions and random sets, it is easy to define belief functions on the real line [Dempster, 1968; Strat, 1984; Smets, 2005; Denœux, 2009], or even on more abstract topological spaces [Shafer, 1973, 1979; Nguyen, 1978, 2006]. We can also extend belief and plausibility functions to fuzzy events [Smets, 1981] by means of Choquet integrals:

$$Bel(F) = \sum_{E \subseteq S} m(E) \cdot \min_{s \in E} F(s) \tag{5}$$

and

$$Pl(F) = \sum_{E \subseteq S} m(E) \cdot \max_{s \in E} F(s), \tag{6}$$

for the finite case. It is also possible to "fuzzify" the theory of belief functions by allowing either the focal sets to be fuzzy sets [Zadeh, 1979; Yen, 1990], or the belief masses to be intervals or fuzzy numbers [Denœux, 1999, 2000a].

Two Special Cases

Two remarkable special kinds of belief functions are worth noticing:

1. Probability functions are obtained by assuming the focal sets to be singletons. It is clear that, if $m(A) > 0$ implies $\exists s \in S, A = \{s\}$, then $Bel(A) = Pl(A) = P(A)$ is the probability function such that $P(\{s\}) = m(\{s\}), \forall s \in S$. Conversely, *Bel* is a probability function if and only of $Bel(A) = Pl(A), \forall A \subseteq S$.
2. Plausibility functions are possibility measures (or, dually, belief functions are necessity measures) if and only of the focal sets are nested, i.e., if $\forall E \neq F \in \mathscr{F}, E \subset F$ or $F \subset E$. In that case, $Pl(A \cup B) = \max(Pl(A), Pl(B))$ and $Bel(A \cap B) = \min(Bel(A), Bel(B))$. For instance, a simple mass function, as in the above watch example, yields possibility and necessity measures.

We can associate to m the mapping $\varphi_m : S \to [0,1]$ called *contour function* of m defined by $\varphi_m(s) = \mathrm{Pl}(\{s\})$, i.e.,

$$\forall s \in S, \quad \varphi_m(s) = \sum_{s \in E} m(E). \tag{7}$$

It is easy to see that function φ_m is normalized in the sense of possibility theory ($\varphi_m(s) = 1$ for some state $s \in S$) whenever the focal sets have a nonempty intersection (which is the case if they are nested). Recovering the mass function m from φ_m is only possible when the focal sets are either nested or disjoint. In particular, if *Bel* is a probability measure, φ_m coincides with m and is a probability distribution. Now assume that the focal sets are nested and form an increasing sequence $E_1 \subset E_2 \subset, \ldots, \subset E_n$, where $E_i = \{s_1, \ldots, s_i\}$; then φ_m is indeed a possibility distribution π, and (7) reduces to $\pi(s_i) = \sum_{j=i}^{n} m(E_j)$. The possibility measure Π and the necessity measure N defined from π coincide, respectively, with the plausibility and belief functions induced by m. The mass function can be recomputed from π as follows (with the notation $\pi(s_{n+1}) = 0$),

$$m_\pi(E_i) = \pi(s_i) - \pi(s_{i-1}), \quad i = 1, \ldots, n. \tag{8}$$

2.3 Combination Rules

The combination of information or evidence from different sources plays a fundamental role in the theory of belief functions. The basic tool is *Dempster's rule of combination* [Dempster, 1967; Shafer, 1976], which makes it possible to combine independent pieces of information. This tool, as well as the precise definition of independence in this context can be introduced using the random code metaphor introduced in Section 2.1.

2.3.1 Dempster's Rule of Combination

Let m_1 and m_2 be two mass functions on S induced by random sets (C_1, P_1, Γ_1) and (C, P_2, Γ_2), where C_1 and C_2 are, as before, interpreted as sets of codes. Assume both codes are selected independently at random. For each pair $(c_1, c_2) \in C_1 \times C_2$, the probability that c_1 and c_2 are jointly selected is $P_1(\{c_1\})P_2(\{c_2\})$; we can then deduce that $v \in \Gamma_1(c_1) \cap \Gamma_2(c_2)$. If moreover we assume the two bodies of evidence pertain to the same message, we have to restrict to cases where $\Gamma_1(c_1) \cap \Gamma_2(c_2) \neq \emptyset$. Consequently, the joint probability distribution on $C_1 \times C_2$ should be conditioned on the set $\{(c_1, c_2) \in C_1 \times C_2 | \Gamma_1(c_1) \cap \Gamma_2(c_2) \neq \emptyset\}$. This line of reasoning leads to the following rule, called Dempster's rule or the product-intersection rule,

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C) \tag{9}$$

for any $A \subseteq S$, $A \neq \emptyset$ and $(m_1 \oplus m_2)(\emptyset) = 0$, where

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{10}$$

is called the *degree of conflict* between m_1 and m_2. If $\kappa = 0$, the two bodies of evidence are said to be *non-conflicting*, i.e., each focal set of m_1 intersects all focal sets of m_2. If $\kappa = 1$, the two bodies of evidence are logically contradictory and, consequently, they cannot be combined. Mass function $m_1 \oplus m_2$ is called the *orthogonal sum* of m_1 and m_2. The unnormalized version of this rule was introduced by Smets [1990a]. A general definition of Dempster's rule in infinite spaces was given by Shafer [1973, 2016a].

Dempster's rule is commutative, associative and it admits the vacuous mass function $m^?$ as neutral element. It can be easily computed using the commonality function (4). Denoting by Q_1, Q_2 and $Q_1 \oplus Q_2$ the commonality functions associated, respectively to m_1, m_2 and $m_1 \oplus m_2$, the following relation holds,

$$Q_1 \oplus Q_2 = \frac{1}{1-\kappa} Q_1 \cdot Q_2. \tag{11}$$

2.3.2 Dempster's Rule of Conditioning

Conditioning in evidence theory, referred to as *Dempster's rule of conditioning*, was proposed by Shafer [1976]. It is a special case of Dempster's rule of combination (cf. Section 2.3.1), mass function m being combined with a logical mass function m_C such that $m_C(C) = 1$. The idea is to transfer all the mass from each focal set E to $E \cap C \neq \emptyset$, since m_C states that the truth lies in C, and to renormalize the obtained result. The new information C can then be viewed as a revision of the original belief function so as to ensure that $Pl(\overline{C}) = 0$: the situations in which C is false are now considered as impossible. Denoting by $Pl(A \parallel C)$ the revised plausibility, we have

$$Pl(A \parallel C) = \frac{Pl(A \cap C)}{Pl(C)}, \tag{12}$$

which clearly constitutes an extension of probabilistic conditioning. The conditional belief function is then obtained dually as $Bel(A \parallel C) = 1 - Pl(\overline{A} \parallel C)$. We can remark that, with this rule of conditioning, the size of focal sets decreases: consequently, information becomes more precise, and the intervals $[Bel(A), Pl(A)]$ become narrower (up to the normalization factor). Especially, when $Bel(C) = 0$ and $Pl(C) = 1$ (total ignorance about C), conditioning on C by Dempster's rule increases the precision of the resulting mass function. Indeed, Dempster's conditioning corresponds to a revision process leading to information enrichment. Revision is here viewed as the combination between a body of uncertain evidence and a sure piece of information.

2.3.3 Other Combination Rules

Dempster's rule tends to concentrate belief masses on smaller focal sets: it thus has a conjunctive behavior. We can define a *disjunctive* counterpart to Dempster's rule [Dubois and Prade, 1986; Smets, 1993] as follows,

$$\forall A \subseteq S, \quad (m_1 \cup m_2)(A) = \sum_{B \cup C = A} m_1(B) m_2(C). \tag{13}$$

This combination rule assumes that at least one of the two information sources is reliable, contrary to Dempster's rule, which assumes that they both are reliable. The disjunctive rule is commutative, associative, and admits as neutral element the mass function m such $m(\emptyset) = 1$. It can be expressed from belief functions using product:

$$(Bel_1 \cup Bel_2) = Bel_1 \cdot Bel_2, \tag{14}$$

which can be compared to (11). Note that the weighted average of belief functions is still a belief function It offers yet another alternative combination rule. The set of belief functions is thus closed under product and weighted average.

2.3.4 Approximations by Reducing the Number of Focal Sets

Both Dempster's rule (9) and its dual disjunctive rule (13) have the effect of increasing the number of focal sets. To avoid combinatorial explosion, a useful strategy is to approximate each mass function by a simpler one with fewer focal sets. Several methods with different degrees of complexity have been proposed for this purpose [Lowrance et al, 1986; Tessem, 1993; Bauer, 1997; Harmanec, 1999; Denœux, 2001]. The simplest, yet quite effective approach, is the *Summarization* algorithm [Lowrance et al, 1986], which works as follows. Let $F_1,\ldots,F_n$ be the focal sets of m ranked by decreasing mass, i.e., $m(F_1)\geq m(F_2)\geq\ldots\geq m(F_n)$. If n exceeds some the maximum allowed number k of focal sets, then the $n-k$ focal sets F_i, $i=k+1,\ldots,n$ with the smallest masses are replaced by their union, and m is approximated by the mass function m' defined as

$$m'(F_i)=m(F_i),\quad i=1,\ldots,k, \tag{15a}$$

$$m'\left(\bigcup_{i=k+1}^{n}F_i\right)=\sum_{i=k+1}^{n}m(F_i). \tag{15b}$$

A more sophisticated algorithm for grouping focal sets while minimizing information loss, based on the principle of hierarchical clustering, was proposed by Denœux [2001].

When Equations (11) or (14) are used, the complexity depends no longer on the number of focal sets, but on the cardinality of the universe S. An efficient approximation algorithm based on the search for a coarsening (grouping of focal sets) minimizing information loss was proposed by Denœux and Ben Yaghlane [2002]. Using a completely different approach, the combination of several belief functions can also be performed by Monte-Carlo simulation (see, e.g., [Moral and Wilson, 1994, 1996]).

2.3.5 Conflict Management

The management of conflict between information sources in an important practical problem, which has drawn a lot of attention over the years [Lefèvre et al, 2002; Smets, 2007; Martin et al, 2008; Destercke and Burger, 2013]. When a high conflict between pieces of information is detected, two strategies are possible: we can either revise the way information has been formalized, or we can use a *robust* combination rule yielding a consistent result in case of conflict.

An example of such rule is the Dubois and Prade [1986] rule defined as follows,

$$(m_1\circledast m_2)(A)=\sum_{B\cap C=A}m_1(B)m_2(C)+\sum_{\{B\cap C=\emptyset,B\cup C=A\}}m_1(B)m_2(C), \tag{16}$$

for any $A\subseteq\Omega$, $A\neq\emptyset$, and $(m_1\circledast m_2)(\emptyset)=0$. When the degree of conflict κ between m_1 and m_2 is zero, we get $m_1\circledast m_2=m_1\oplus m_2$: in the absence of conflict, the

Dubois-Prade rule is equivalent to Dempster's rule. In contrast, when the degree of conflict is equal to 1, we have $m_1 \circledast m_2 = m_1 \cup m_2$: in that case, the Dubois-Prade rule boils down to the disjunctive rule. In all other cases, the behavior of the $\circledast$ operator is intermediate between conjunctive and disjunctive modes: it is an *adaptive* combination rule. We can remark that this rule is commutative but it is not associative. However, an n-ary version can easily be defined, based on maximal consistent subsets of focal sets [Dubois and Prade, 1986]. More complex ways of distributing the conflict among focal sets have been proposed (see, e.g., [Lefèvre et al, 2002; Martin et al, 2008]). See also Chapter 14 in this volume, for more details on fusion operations.

2.3.6 Combination of Dependent Information

Dempster's rule (9) and its disjunctive counterpart (13) both make an independence assumption about the pieces of information to be combined. While it is often possible to break down a body of evidence into independent pieces [Shafer, 2016c], this is not always the case, especially in sensor fusion applications. It is then useful to have a well-justified rule allowing us to combine non independent pieces of evidence.

Such a rule, called the *cautious rule*, was proposed by Denœux [2008]. It is based on the weight function representation, which we will now introduce. A mass function m is said to be *separable* [Shafer, 1976] if it the orthogonal sum of simple mass functions (see Section 2.1). Denoting a simple mass function with focal sets A and S as $A^{w(A)}$, where $w(A)$ is the mass committed to S (so, $1-w(A)$ is committed to A), a separable mass function can thus be written as

$$m = \bigoplus_{\emptyset \neq A \subset S} A^{w(A)}. \tag{17}$$

Considering the negation $\overline{m}$ of a mass function m, defined by $\forall A, \overline{m} = m(\overline{A})$ [Dubois and Prade, 1986], there is a De Morgan duality between the disjunctive rule (14) and the non-normalized variant of Dempster's rule (9) that has been exploited by Denœux [2008] to define a disjunctive decomposition of belief functions.

Given a separable mass function m with commonality function Q such that $m(S) > 0$, the weights $w(A)$ can be recovered from Q as

$$\ln w(A) = -\sum_{B \supseteq A} (-1)^{|B|-|A|} \ln Q(B), \quad \forall A \subset S, A \neq \emptyset. \tag{18}$$

The mapping $w : 2^S \setminus \{\emptyset, S\} \rightarrow [0,1]$ defined by (18) is called the *weight function* associated to m. When m is not separable but still verifies $m(S) > 0$ (it is then said to be *non dogmatic*), we can still define the weight function w from (18), but we can now have $w(A) > 1$ for some A [Smets, 1995]. Mass function m can then still computed from w using (17), where $A^{w(A)}$ with $w(A) > 1$ is no longer a proper mass

function but a *generalized mass function* assigning "masses" $w(A) > 1$ to S and $1 - w(A) < 0$ to A.

Given two non dogmatic mass function m_1 and m_2 with weight functions w_1 and w_2, their orthogonal sum can be written as:

$$m_1 \oplus m_2 = \bigoplus_{\emptyset \neq A \subset \Omega} A^{w_1(A) w_2(A)},$$

i.e., the weight function of $m_1 \oplus m_2$ is the product of those of m_1 and m_2. In contrast, the cautious rule is defined as

$$m_1 \circledcirc m_2 = \bigoplus_{\emptyset \neq A \subset \Omega} A^{\min(w_1(A), w_2(A))}, \tag{19}$$

i.e., the weight function of $m_1 \circledcirc m_2$ is the minimum of those of m_1 and m_2. The cautious rule is commutative, associative *and idempotent*, which makes it suitable to combine dependent pieces of evidence. It can be justified by the Least Commitment Principle (see Section 2.4). A disjunctive counterpart of $\circledcirc$, called the bold disjunctive rule, can also be defined [Denœux, 2008]. With Dempster's rule and the disjunctive rule (13), the cautious and bold rules can be seen as particular elements of infinite families of rules based on triangular norms and on uninorms [Pichon and Denœux, 2010]. Other idempotent, but non associative rules have been defined and studied by Destercke and Dubois [2011] and Cattaneo [2011].

2.3.7 Taking Into Account Metaknowledge About Sources

When combining information from several sources, it is often useful to take into account not only the information provided by the sources, but also metaknowledge about their properties (such as their reliability or truthfulness). The *discounting* operation [Shafer, 1976; Smets, 1993] makes it possible to account for the reliability of a source by transferring a fraction α of each mass $m(A)$ for $A \subset S$ to S. The discounted mass function, denoted by ${}^{\alpha}m$, is then given by

$${}^{\alpha}m = (1-\alpha)\, m + \alpha\, m^{?},$$

where, as before $m^{?}$ denotes the vacuous mass function and α is called the *discount rate*. The *contextual discounting* operation, introduced by Mercier et al [2008], generalizes discounting by allowing one to take into account the source's reliability in different contexts. Pichon et al [2012] have proposed a very general mechanism for "correcting" and combining mass functions, taking into account the relevance and truthfulness of information sources; they have shown that all connectives of Boolean logic can be interpreted in the light of these two properties. Other belief function correction mechanisms have been proposed by Mercier et al [2012, 2016], and Pichon et al [2016].

2.4 Imprecision, Specialization and Information Measures

Like any information items, it is interesting to compare belief functions according to their information content. This makes it possible, in particular, to apply the *maximum uncertainty* ([Klir and Wierman, 1999]) or *least commitment* ([Smets, 1993]) principle, which serves the same purpose as the maximum entropy principle in probability theory and the principle of minimal specificity in possibility theory. According to this principle, when several belief functions are compatible with a set of constraints, the least committed should be selected. In order to apply this principle, we need to define a partial order on the set of belief functions. For that purpose, we may either define a degree of imprecision or of uncertainty of a belief function, or we may adopt a more qualitative approach and directly define an informational ordering relation on the set of belief functions.

2.4.1 Quantitative Approach

As belief functions model both imprecise and uncertain information, we may be willing to measure imprecision and uncertainty separately. A natural measure of imprecision is the *expected cardinality* of the random set defined by the mass function,

$$\mathrm{Imp}(m) = \sum_{E \subseteq S} m(E) \cdot \mathrm{card}(E). \tag{20}$$

It is clear that $\mathrm{Imp}(m^?) = \mathrm{card}(S)$, where $m^?$ is the vacuous mass function, and $\mathrm{Imp}(m) = 1$ when m is a probability mass function. It can be checked that $\mathrm{Imp}(m) = \sum_{s \in S} Pl(\{s\})$. An alternative measure of imprecision is *nonspecificity* [Dubois and Prade, 1985], defined for a normalized mass function m as

$$N(m) = \sum_{E \subseteq S} m(E) \log_2 \mathrm{card}(E). \tag{21}$$

Nonspecificity was shown by Ramer [1987] to be the only measure of imprecision satisfying some rationality requirements.

The degree of uncertainty of a belief function can be measured by generalizing the well-known Shannon entropy of a probability measure P defined by

$$H(P) = -\sum_{i=1}^{\mathrm{card}(S)} p_i \cdot \ln p_i. \tag{22}$$

Several extensions of $H(p)$ to belief functions have been proposed, of the form

$$D(m) = -\sum_{E \subseteq S} m(E) \cdot \ln g(E), \tag{23}$$

where g can be, e.g., Pl or Bel [Dubois and Prade, 1987; Klir and Wierman, 1999]. For $g = Pl$, we get a measure of *dissonance* (or internal conflict), which is maxi-

mized by uniform probability measures, and reaches its minimum (zero) when the intersection of focal sets is non empty : $\bigcap\{E : m(E) > 0\} \neq \emptyset$. For $g = Bel$, we rather have a measure of *confusion*, which is minimal (zero) for logical mass functions verifying $m(E) = 1$ for some unique focal set E (imprecise but certain and clear information), but high for uniform mass functions over subsets of S with cardinality $\text{Card}(S)/2$ [Dubois and Ramer, 1993]. See also [Ramer and Klir, 1993; Klir and Wierman, 1999].

Another approach, proposed by Smets [1983], is to define a measure I of information content that relies on the pivotal role of Dempster's rule in the theory of belief functions, namely, it is natural to impose an additivity property with respect to this rule, such as $I(m_1 \oplus m_2) = I(m_1) + I(m_2)$ for any two non-conflicting mass functions m_1 and m_2. As shown by Smets [1983], this requirement, together with a few additional natural conditions, lead to the following definition: [1]

$$I(m) = -\sum_{E \subseteq S} \ln Q(E). \tag{24}$$

Other quantitative criteria attempt to measure imprecision and uncertainty simultaneously. For instance, *aggregate uncertainty* [Maeda and Ichihashi, 1993; Harmanec and Klir, 1994] is defined as follows, for a normalized mass function m:

$$AU(m) = \max_{P \in \mathscr{P}(m)} H(P), \tag{25}$$

where H is the Shannon entropy, and $\mathscr{P}(m)$ is the set of probability measures on S compatible with m:

$$\mathscr{P}(m) = \{P, P(A) \leq Pl(A), \forall A \subseteq S\}. \tag{26}$$

It is clear that $AU(m)$ is maximal both for the vacuous mass function $m = m^?$ and for the uniform Bayesian mass function m such that $m(\{s\}) = 1/\text{card}(S)$ for all $s \in S$; these two mass functions correspond, respectively, to maximal imprecision and to maximal uncertainty. Aggregate uncertainty can be shown to meet a number of reasonable requirements [Klir and Wierman, 1999]. However, the debate on what should be a "natural" measure of total uncertainty in the theory of belief functions is not settled: see, for instance, the recent proposal and discussion by Jiroušek and Shenoy [2016].

2.4.2 Comparative Approach

The second approach to comparing the informational contents of belief functions consists of directly defining a partial order on the set of belief functions. Given two normalized mass functions m_1 and m_2, m_1 is said to be *more precise* than m_2

[1] Considering the disjunctive rule instead of the conjunctive rule would lead to replace Q by Bel in (24).

(denoted by $m_1 \sqsubseteq_{Pl} m_2$) iff, for any subset A of S, the interval $[Bel_1(A), Pl_1(A)]$ is included in the interval $[Bel_2(A), Pl_2(A)]$. Because of the duality of Bel and Pl, this condition can be simplified to: $\forall A, Pl_1(A) \leq Pl_2(A)$. In term of imprecise probabilities, the condition $m_1 \sqsubseteq_{Pl} m_2$ means that $\mathscr{P}(m_1)$ is a subset of $\mathscr{P}(m_2)$ [Dubois and Prade, 1986; Yager, 1986]. Mass function m is thus maximally precise when it coincides with a single probability measure, and minimally precise if $m = m^?$. It is also clear that, if $m_1 \sqsubseteq_{Pl} m_2$, then $AU(m_1) \leq AU(m_2)$. Note that this approach is in agreement with the imprecise probability interpretation of belief functions.

An alternative method for comparing the informativeness of belief functions consists in generalizing relative specificity, viewed as set inclusion, to random sets. A normalized mass function m_1 is a *specialization* of a normalized mass function m_2 (denoted by $m_1 \sqsubseteq_s m_2$) if and only of the following three conditions hold:

1. Any focal set of m_2 contains at least one focal set of m_1;
2. Any focal set of m_1 is included in at least one focal set of m_2;
3. There exists a stochastic matrix W whose element w_{ij} is the proportion of the mass $m_1(E_i)$ assigned to $F_j \supseteq E_i$ in order to reconstruct mass $m_2(F_j)$, i.e., $m_2(F_j) = \sum_i w_{ij} \cdot m_1(E_i)$.

This relation is stronger than the previous one: if m_1 is a specialization of m_2, then m_1 is also more precise than m_2 (but the converse is not true in general, see [Dubois and Prade, 1986]). It is also obvious that, if m_1 is specialization of m_2, then $\text{Imp}(m_1) \leq \text{Imp}(m_2)$.

As noted in Section 2.2, in the consonant case, m_π and π contain the same information, i.e., $Pl = \Pi$ and $Bel = N$. Accordingly, for possibility measures, the precision and specialization orderings both coincide with the specificity ordering for possibility distributions: m_{π_1} is a specialization of m_{π_2} iff $\Pi_1(A) \leq \Pi_2(A), \forall A \subseteq S$ iff $\pi_1(s) \leq \pi_2(s), \forall s \in S$ [Dubois and Prade, 1986].

Other informational orderings have been proposed. For instance, m_1 is said to be more informative than m_2 according to commonalities (denoted by $m_1 \sqsubseteq_Q m_2$) iff $Q_1 \leq Q_2$ [Dubois and Prade, 1986; Yager, 1986]. This property can be interpreted from Eq. (11): as numbers $Q_1(A)$ get closer to 1, the influence of m_1 when combined by Dempster's rule with another mass function m_2 becomes smaller, which means that m_1 becomes less informative. Relation $\sqsubseteq_Q$ is weaker than $\sqsubseteq_s$, but it is not comparable with $\sqsubseteq_{Pl}$. Obviously, $m_1 \sqsubseteq_Q m_2$ implies that $I(m_1) \geq I(m_2)$.

Yet another ordering relation was proposed by Denœux [2008], based on the weight function (18). Mass function m_1 is said to be more informative than m_2 according to the weights (denoted by $m_1 \sqsubseteq_w m_2$) iff $w_1 \leq w_2$. This means that m_1 is the orthogonal sum of m_2 and a separable mass function m that has no conflict with m_2: $m_1 = m_2 \oplus m$. The cautious rule (19) can be derived from the least commitment principle based on relation $\sqsubseteq_w$.

2.5 Criteria for Decision Under Uncertainty

Consider a set $\mathscr{A} = \{a_1, \ldots, a_r\}$ of acts, a set $S = \{s_1, \ldots, s_n\}$ of states of nature, and a payoff matrix U of size $r \times n$, whose element u_{ij} is the utility of choosing act a_i if state s_j occurs. Assuming the uncertainty about the state of nature to be modeled by a mass function m on S, which act should be chosen? To answer this question, the classical Maximum Expected Utility (MEU) principle [von Neumann and Morgenstern, 1944] can be generalized in a number of ways in the belief function setting (see also Chapter 17 in this volume).

2.5.1 Lower and Upper Expected Utilities

According to Dempster [1967] and Shafer [1981], the *lower* and *upper expected utilities* of act a_i are defined, respectively, as the following Choquet integrals (further studied in Chapter 17 of this volume) similar to (5):

$$\underline{\mathrm{EU}}(a_i) = \sum_{E \subseteq S} m(E) \min_{s_j \in E} u_{ij} \tag{27a}$$

$$\overline{\mathrm{EU}}(a_i) = \sum_{E \subseteq S} m(E) \max_{s_j \in E} u_{ij}. \tag{27b}$$

The lower and upper expected utilities can be shown to be, respectively, the lower and upper bounds of the expected utility with respect to all probability measures P on S compatible with m [Shafer, 1981]. An optimistic decision-maker (DM) will typically maximize the upper expected utility, while a pessimistic DM will maximize the lower expected utility. These two decision rules can be generalized by considering a convex sum of the lower and upper expected utility [Jaffray, 1989; Strat, 1990], which generalizes Hurwicz criterion (see Chapter 17 in this volume for a detailed discussion of its aximatization due to Jaffray):

$$\mathrm{EU}_\alpha(a_i) = \sum_{E \subseteq S} m(E) \left(\alpha \min_{s_j \in E} u_{ij} + (1-\alpha) \max_{s_j \in E} u_{ij} \right) \tag{28a}$$

$$= \alpha \underline{\mathrm{EU}}(a_i) + (1-\alpha) \overline{\mathrm{EU}}(a_i), \tag{28b}$$

where α can be seen as a pessimism index. An even more general approach, proposed by Yager [1992], combines the utilities in each set $\{u_{ij} \mid s_j \in E\}$ by an Ordered Weighted Average (OWA) operator.

2.5.2 Pignistic Probability

Following a different line of reasoning and putting emphasis on the avoidance of Dutch books (i.e., sequences of bets ensuring a sure loss), Smets [1990b] advocated a two-level mental model: the *credal* level, where uncertainty is represented

by a belief function, and the *pignistic* level, where belief functions are transformed to probabilities for decision-making. The *pignistic transformation* [Smets, 1990b] consists in distributing each mass $m(E)$ equally to all elements of E, resulting in the probability distribution *betp* defined as

$$betp(s) = \sum_{E:s\in E} \frac{m(E)}{\text{card}(E)}. \tag{29}$$

This transformation had been earlier proposed by Dubois and Prade [1982] as a generalization of Laplace's principle of insufficient reason to belief functions. Smets [1990b] justified it axiomatically, by imposing a linearity property (the pignistic probability of a convex sum of belief functions should be the convex sum of the pignistic probabilities) and an anonymity property (the pignistic probability of an event E should not change after permuting the elements of E). In fact, the pignistic probability was already known in the theory of cooperative games since the 1950's as the *Shapley value*, and Smets' axioms are mathematically the same as those proposed by Shapley [1953], albeit in a different context. The pignistic probability is also the center of gravity of the convex set of probabilities that dominate the belief function.

We can also search for the least informative belief function, according to the commonality ordering $\sqsubseteq_Q$ defined in Section 2.4.2, corresponding to a given pignistic probability distribution. As shown by Dubois et al [2008], it is unique and consonant; consequently, it induces a possibility distribution.

Having defined the pignistic distribution *betp*, we can evaluate each act a_i by its expected utility with respect to *betp*,

$$\text{EU}_{betp}(a_i) = \sum_{s_j\in S} betp(s_j)u_{ij} = \sum_{E\subseteq S} m(E)\left(\frac{1}{\text{card}(E)}\sum_{s_j\in E} u_{ij}\right), \tag{30}$$

which can be compared to (27) and (28a). The pignistic criterion is a special case of Yager's OWA criterion [Yager, 1992], as the average is a particular OWA operator.

2.6 Applications to Statistical Learning and Data Analysis

In Artificial Intelligence, the theory of belief functions has been used, until the early 1990's, to model uncertainty in expert systems [Shafer, 1987; Shenoy, 1989]. Since the 1990, we have seen the development of another application area: statistical learning (see Chapter 12 of Volume 2). The theory of belief functions has proved to be an efficient formalism for combining models, modeling uncertainty in the outputs of classifiers or clustering algorithms, and learning from uncertain data. In the following, we review some of the recent developments in this area.

2.6.1 Classifier Combination

A first way of applying the theory of belief functions to classification is to consider classifier outputs as items of evidence and to merge them using Dempster's rule, or any other combination rule (see Section 2.3). Given the flexibility of the belief function formalism, this approach can be applied to classifiers of various types, the outputs of with can be converted into belief functions.

For instance, Xu et al [1992] proposed to use a confusion matrix to convert a classifier's decision into a mass function. They obtained good results for a handwriting recognition problem. A similar approach was used by Mercier et al [2009] for postal address recognition. More recently, Bi et al [2008] proposed to represent classifier scores as "triplet" mass functions with three focal sets. Bi [2012] studied the influence of classifier diversity and the combination rule on the accuracy of the ensemble. Quost et al [2011] considered a parametrized family of combination rules, including Dempster's rule and the cautious rule (see Section 2.3.6), and proposed a method to find the best rule in this family.

From a different perspective, Quost et al [2007] considered the combination of binary classifiers as a way to solve multi-class classification problems. For instance, in the "one-against-one" approach, binary classifiers are trained using data from only two classes; consequently, their outputs can be interpreted as conditional mass functions. The problem is then to construct an unconditional mass function on the whole set of classes, as consistent as possible with the conditional mass functions provided by the binary classifiers.

2.6.2 Evidential Classifiers

An *evidential classifier* is a classifier whose output is a mass function over a set of classes $\Omega = \{\omega_1, \ldots, \omega_c\}$. Two main approaches have been proposed for constructing such a classifier from training data.

The first approach, first introduced and justified axiomatically by Appriou [1991, 1998], is to construct a mass function m on Ω from the likelihoods $p(x|\omega_k)$, where x is the feature vector. One of the two methods proposed by Appriou is identical to the solution resulting from the application the *Generalized Bayes Theorem* (GBT) introduced by Smets [1993]. The mass function has the following expression:

$$m = \bigoplus_{k=1}^{c} \overline{\{\omega_k\}}^{\alpha_k p(x|\omega_k)}, \tag{31}$$

where the α_k's are coefficients ensuring that $\alpha_k p(x|\omega_k) \leq 1$ for $k = 1, \ldots, c$, and the notation A^w stands for the simple mass function μ such that $\mu(A) = 1 - w$ and $\mu(\Omega) = w$ (see Section 2.3.6). A major advantage of this method is that it can be used without prior class probabilities, or with only weak prior information encoded as a belief function. However, when prior probabilities are given, the GBT yields the same solution as the Bayesian approach. Appriou [1991] showed the robustness

of this method, in particular when the test data distribution differs from the learning distribution due, e.g., to different data acquisition methods or to sensor malfunction.

Another approach, referred to as the *evidential k-nearest neighbor (NN) rule*, was introduced by Denœux [1995]. It consists in considering each training instance (or only each of the k nearest instances in the training set) as a piece of evidence about the class of the new object to be classified. The different pieces of evidence are represented by mass functions and are combined using Dempster's rule. In the most general form of this method, we consider a training set

$$\mathscr{L} = \{(x^{(1)}, m^{(1)}), \ldots, (x^{(N)}, m^{(N)}),$$

where $x^{(i)}$ is the feature vector of instance i and $m^{(i)}$ is a mass function on Ω representing partial knowledge about the class of that example. In the fully supervised case, each mass function $m^{(i)}$ is certain, i.e., we have $m^{(i)}(\{\omega_j\}) = 1$ for some element ω_j of Ω. In the general case, we have a *partially supervised* learning problem. Partial knowledge about the class of training instances may be provided by an expert or derived from indirect observation. We also assume a distance or dissimilarity measure δ between feature vectors.

The mass function representing the evidence of the training example $e^{(i)} = (x^{(i)}, m^{(i)})$ is defined as

$$m(A \mid e^{(i)}) = \varphi\left(\delta(x, x^{(i)})\right) m^{(i)}(A), \quad \forall A \subset \Omega \tag{32a}$$

$$m(\Omega \mid e^{(i)}) = 1 - \sum_{A \subset \Omega} m(A|e^{(i)}), \tag{32b}$$

where φ is a decreasing function verifying $\varphi(0) \leq 1$ and $\lim_{d\to\infty} \varphi(d) = 0$. Mass function $m(\cdot|e^{(i)})$ is thus obtained by discounting $m^{(i)}$ (see Section 2.3.7), with a discount rate that gets closer to one when the dissimilarity between vectors x and $x^{(i)}$ goes to infinity. The condition $\lim_{d\to\infty} \varphi(d) = 0$ ensures that mass function $m(\cdot|e^{(i)})$ becomes vacuous when the dissimilarity between vectors x and $x^{(i)}$ goes to infinity.

Let us now consider a new object described by a known feature vector $\hat{x}$ and an unknown class label $y \in \Omega$. Having computed mass functions (32) for each of the K nearest neighbors of $\hat{x}$, the combined mass function on Ω is

$$m(\cdot \mid \mathscr{L}) = \bigoplus_{\{i|x_i \in \mathscr{N}_K(\hat{x})\}} m(\cdot \mid e^{(i)}), \tag{33}$$

where $\mathscr{N}_K(\hat{x})$ denotes the set of the K nearest neighbors of $\hat{x}$. The choice of a best class $\hat{y} \in \Omega$ can then be made using one of the decision rules described in Section 2.5 and in Chapter 17 of this volume. Denœux [1997] describes several decision strategies with different reject options.

Zouhal and Denœux [1998] have proposed a method for choosing function φ within a parametric family by minimizing an error function. The *evidential neural network classifier* introduced by Denœux [2000b] is a variant of this method, in which the training set is summarized as a set of prototypes. Both the evidential k-

NN rule and the evidential neural network classifier have been implemented in the R package `evclass` [Denoeux, 2017]. Denœux and Zouhal [2001] have studied another variant of the evidential k-NN rule in which partial information about the class of training instances is given as possibility distributions. Petit-Renaud and Denœux [2004] have extended the approach to regression problems, where variable y is numerical. Recently, Lian et al [2015] proposed a feature selection method based on the evidential k-NN rule, and Lian et al [2016] described an algorithm for learning the distance function δ in (32).

The evidential k-NN rule has also been extended to multi-label classification problems, in which each object may belong simultaneously to several classes [Denoeux et al, 2010]. In this case, the universe is the power set 2^Ω of the set of classes. To prevent double exponential complexity in the manipulation of mass functions, belief functions can then be defined on a lattice of subsets of Ω (the intervals with respect to the ordering relation $\subseteq$). A general presentation of this approach (with applications not only to classification, but also to preference elicitation and to clustering) can be found in [Denœux and Masson, 2012]. See also [Grabisch, 2009] for the general theory of belief functions on lattices.

The likelihood-based and distance-based evidential classification methods outlined above have been compared experimentally by Fabre et al [2001], and theoretically by Denœux and Smets [2006], who showed that they can both be derived from the GBT.

2.6.3 Evidential Clustering

The theory of belief functions has also been applied to clustering, which consists in finding groups (or clusters) in data (see Chapters 12 and 14 of Volume 2). Here, belief functions can be used to quantify the uncertainty about the group membership of each particular object. Given a set of n objects $\mathscr{O} = \{o_1, \ldots, o_n\}$ and a set of c clusters $\Omega = \{\omega_1, \ldots, \omega_c\}$, Denœux and Masson [2004] defined a *credal partition* as an n-tuple $M = (m_1, \ldots, m_n)$ of (not necessarily normalized) mass functions on Ω, where m_i quantifies the uncertainty about the cluster membership of object o_i. A credal partition boils down to a hard partition when all mass functions are precise (i.e., when they focus on only one singleton). Most other "soft" clustering notions such as fuzzy, possibility and rough clustering are also recovered as special cases [Denoeux and Kanjanatarakul, 2016]. For instance, if all mass functions correspond to probability distributions (i.e., their focal sets are singletons), then we can identify each mass $m_i(\{\omega_k\})$ with the degree of membership u_{ik} of object o_i to cluster ω_k, and we have a fuzzy partition [Bezdek, 1981]. If each mass function m_i is categorical (i.e., it has only one focal set A_i), then we can define the *lower approximation* of cluster ω_k as the set of objects o_i that surely belong to ω_k, i.e., such that $A_i = \{\omega_k\}$, and the *upper approximation* of cluster ω_k as the set of objects o_i that may belong to ω_k, i.e., such that $\omega_k \in A_i$. We then have a *rough partition* as defined by Lingras and Peters [2012]. A general credal partition can also easily be summarized into a hard partition or any type of soft partition. For instance, we obtain a fuzzy partition

by replacing each mass m_i by its pignistic probability distribution (29), and we get a rough partition by selecting, for each mass function m_i, the focal set with the largest mass [Denoeux and Kanjanatarakul, 2016].

An *evidential clustering* algorithm is a procedure that constructs a credal partition from a dataset. Several such algorithms have been proposed over the years:

- The *EVCLUS* algorithm, introduced in [Denœux and Masson, 2004], applies ideas from multidimensional scaling to clustering: given a dissimilarity matrix, it finds a credal partition such that the degrees of conflict (10) between mass functions match the dissimilarities, dissimilar objects being represented by highly conflicting mass functions; this is achieved by iteratively minimizing a stress function. A variant of EVCLUS allowing one to use prior knowledge in the form of pairwise constraints was later introduced in [Antoine et al, 2014], and several improvements to the original algorithm making it capable of handling large dissimilarity datasets have been reported in [Denœux et al, 2016] and [Li et al, 2018].
- The *Evidential c-means* (ECM) algorithm [Masson and Denœux, 2008] is a *c*-means-like algorithm that minimizes a cost function by searching alternatively the space of prototypes and the space of credal partitions. Unlike the hard and fuzzy *c*-means algorithms, ECM associates a prototype not only to each cluster, but also to each nonempty set of clusters. The prototype associated to a set of clusters is defined as the barycenter of the prototypes of each single cluster in the set. The cost function to be minimized insures that objects close to a prototype have a high mass assigned to the corresponding set of clusters. A variant with adaptive metrics and binary constraints was introduced in [Antoine et al, 2012], and a relational version for dissimilarity data (called RECM) has been proposed in [Masson and Denœux, 2009]. A version of ECM taking into account spatial constraints and suitable for image segmentation was introduced by Lelandais et al [2014].
- The E*k*-NNclus algorithm [Denœux et al, 2015] is a decision-directed clustering procedure based on the evidential *k*-NN rule described in Section 2.6.2. Starting from an initial partition, the algorithm iteratively reassigns objects to clusters using the evidential *k*-NN rule, until a stable partition is obtained. After convergence, the cluster membership of each object is described by a mass function on Ω assigning a mass to each cluster and to the whole set of clusters. The mass assigned to the set of clusters can be used to identify outliers. The procedure can be seen as searching for the most plausible partition of the data.

All these algorithms have been implemented in the R package `evclust` [Denœux, 2016].

3 Imprecise Probabilities

Imprecise probability theory [Walley, 1991] relies on an approach opposite to the one of belief functions. Instead of randomizing the set-based approach to incomplete information, incompleteness is injected in probability theory. Under the frequentist view, epistemic uncertainty goes on top of a probabilistic model. Under the subjectivist view, the betting protocol is relaxed, by no longer enforcing the equality between buying and selling prices. In the area of economics, Gilboa and Schmeidler [1989] already showed that by suitably relaxing Savage axioms for decision under uncertainty, it is possible to formally justify the idea that an agent's epistemic state consists of a set of probability distributions on the set S of possible states of the world: in order to hedge against uncertainty, when evaluating a decision, the cautious agent picks the probability distribution that minimizes its expected utility.

3.1 Basic Definitions and Interpretations

An imprecise probability model comes down to specifying a family $\mathscr{P}$ of probability functions over S. However, there are several approaches to come up with this family according to the understanding of probability (frequentist or subjectivist), and to the available data in the specific application context.

3.1.1 Incomplete Information About Frequentist Probability

Under a frequentist view, $\mathscr{P}$ is an epistemic set reflecting incomplete information about an otherwise precise mathematical model of a random process: a probability distribution in $\mathscr{P}$ is the right one. The family $\mathscr{P}$ thus represents an imprecise probabilistic model. There are several situations that lead to such a model:

- The most common situation is when several probability measures are compatible with the available information, for instance in the case of scarce data. In the parametric case, the parameters of the model are ill-known, because the confidence intervals for these parameters are too wide. Bayesians then often assume a prior probability distribution on the parameter range or the set of possible probability functions. This is precisely what is not assumed in the imprecise probability setting. Some authors may still use the Bayesian paradigm, but assume imprecision about the prior probability (they are called robust Bayesians [Huber, 1981; Berger, 1994]), resulting in an imprecise posterior distribution.
- Imprecise information can be obtained by an expert or from empirical data about statistical parameters (like support, mean, mode, median, some fractiles) but the type of probabilistic model is otherwise ill-known [Baudrit and Dubois, 2006] (e.g., you know the empirical mean and variance but you do not know if the process is Gaussian or not). It may be that the expert provides probability bounds

on some events (intervals, quantiles, etc.). In the finite case, an expert may assign a probability interval to each outcome instead of a precise value [de Campos et al, 1994].

- A usual setting for getting upper and lower probabilities is the one of imprecise statistical information, that corresponds to Dempster [1967]'s setting for belief functions. The mass value of a focal set is the frequency of observing this incomplete information item. In that case, belief and plausibility functions are lower and upper probabilities, respectively, with a frequentist flavor. See the book by Couso et al [2014] for a presentation of this approach to imprecise statistics.
- Some authors have even questioned the basic assumptions that frequencies converge toward limit probabilities. For instance it is only known that frequencies eventually remain inside an interval [Walley and Fine, 1982].

Suppose one comes up to a probability family $\mathscr{P}$ via some of the above scenarii. Then one can assign to each event lower and upper bounds for the probability of this event [Smith, 1961]:

$$P_*(A) = \inf_{P\in\mathscr{P}} P(A);\ \ P^*(A) = \sup_{P\in\mathscr{P}} P(A). \tag{34}$$

Functions P_* and P^* are monotonic with respect to inclusion and satisfy the duality property $P^*(A) = 1 - P_*(\overline{A})$. We call set functions P_* and P^* *lower and upper envelopes* respectively, after [Walley, 1991]. The additivity property of P enforces the following conditions for such envelopes [Good, 1962]: $\forall A, B \subseteq S$, such that $A \cap B = \emptyset$,

$$P_*(A) + P_*(B) \leq P_*(A \cup B) \leq P_*(A) + P^*(B) \leq P^*(A \cup B) \leq P^*(A) + P^*(B). \tag{35}$$

The width of the interval $[P_*(A), P^*(A)]$ represents the amount of ignorance of the agent as to the truth of proposition A. Total ignorance is when this interval is $[0,1]$. When this interval reduces to a singleton, full probabilistic knowledge is obtained. Probability envelopes are more general than belief and plausibility functions, hence more general than necessity and possibility measures [Walley, 1996].

It is important to notice that in general, it is impossible to reconstruct the original set $\mathscr{P}$ from the knowledge of these intervals $[P_*(A), P^*(A)]$ for all events A. Indeed, these intervals correspond to particular projections of $\mathscr{P}$. Namely, let $\mathscr{P}(P_*) = \{P : \forall A \subseteq S, P(A) \geq P_*(A)\}$, it is easy to see that $\mathscr{P}(P_*)$ is convex (if $P_1 \in \mathscr{P}(P_*)$ and $P_2 \in \mathscr{P}(P_*)$ then, $\forall \lambda \in [0,1], \lambda \cdot P_1 + (1-\lambda) \cdot P_2 \in \mathscr{P}(P_*)$) and contains the convex hull of $\mathscr{P}$ even if $\mathscr{P}$ and $\mathscr{P}(P_*)$ have the same upper and lower envelopes.

A characteristic property of an upper envelope (induced by a non-empty set of probabilities) was found by Giles [1982]. Viewing a set A as its $\{0,1\}$-valued characteristic function ($A(s) = 1$ if $s \in A$ and 0 otherwise). A set-function g is an upper envelope if and only if for any tuple $A_0, A_1, \ldots, A_k$ of subsets of S, and any pair of positive integers (r,s) such that $\sum_{i=1}^{k} A_i(\cdot) \geq r + s \cdot A_0(\cdot)$, it holds that

$$\sum_{i=1}^{k} g(A_i) \geq r + s \cdot g(A_0). \tag{36}$$

3.1.2 The Subjectivist Point of View

The subjectivist approach to imprecise probability was fully developed by Walley [1991]. It is powerful enough to encompass all convex sets of probabilities. In this approach the agent proposes buying prices for gambles. A gamble is a function f from S to the real line that expresses losses ($f(s) < 0$) or gains ($f(s) > 0$). The gamble associated to an event is its characteristic function. The agent is not committed to selling such gambles at the same prices as the ones he or she accepts to buy them.

Informally, the approach relies on so-called *desirable gambles* [Walley, 1991] that the agent would agree to buy for a positive price. The set of desirable gambles contains at least all positive gambles. Moreover the sum of two desirable gambles is desirable, and a desirable gamble remains desirable when multiplied by a positive constant. The lower prevision $LP(f)$ of a gamble f is the maximal value α such that $f - \alpha$ is desirable. It can be shown that given a set of gambles $f_i \in \mathscr{G}$ and their lower previsions $LP(f_i)$, there is a convex set of probabilities $\mathscr{P}$, called *credal set*, such that $LP(f_i)$ is the lower expectation of f_i according to $\mathscr{P}$, for all $f_i \in \mathscr{G}$. One important point is that any convex set of probabilities can be represented by lower previsions on some family of gambles.

In this setting, the upper prevision $UP(f)$ of a gamble f is provably equal to $-LP(-f)$. The value $LP(f)$ is thus the maximal buying price for a gamble f, and the upper prevision $UP(f)(\geq LP(f))$ is the minimal selling price of f. If the credal set attached to a set of gambles and its lower previsions is empty, then the proposal is inconsistent and the agent incurs a sure loss after buying and resolving these gambles. Moreover, due to the interaction between gambles, it may be that the consistent buying prices proposed by the agent for gambles $f_i \in \mathscr{G}$ are too low and could be raised without altering the credal set. A set of buying prices $pr(f_i), f_i \in \mathscr{G}$ is said to be *coherent* if and only if $LP(f_i) = pr(f_i), \forall f_i \in \mathscr{G}$. In other words, letting $E_P(f)$ be the expectation of f with respect to probability P, a set of buying prices for a set of gambles $\mathscr{G}$ is coherent if and only if for any $f_i \in \mathscr{G}, \inf\{E_P(f_i) : P \in \mathscr{P}\} = pr(f_i)$, where $\mathscr{P}$ is the credal set induced by the gambles $f_i \in \mathscr{G}$, and their buying prices. Clearly, Giles condition (36) is easily interpreted in terms of coherence of gambles. It expresses the coherence of a set of upper probabilities assigned to subsets of S (minimal selling prices of 0-1 gambles), protecting an agent who sells $k+1$ lottery tickets corresponding to events $A_0, A_1, \ldots, A_k$ from losing money while proposing optimal selling prices $g(A_i)$.

The gamble approach leads to a decision rule that is specific to the imprecise probability setting, namely a gamble f is preferred to a gamble g if and only the gamble $h = f - g$ is desirable, i.e., if the lower expectation of the latter gamble with respect to the corresponding credal set $\mathscr{P}$ is positive. It gives a partial ordering on gambles. It implies that $\forall P \in \mathscr{P}, E_P(f) \geq E_P(g)$. See Chapter 17 in this volume for other decision rules with credal sets

3.1.3 Special Cases

A monotonic set-function $g : 2^S \to [0,1]$ is said to be a Walley-coherent lower probability if the following property holds:

$$g(A) = \inf\{P(A) : P(A) \geq g(A), \forall A \subseteq S\}.$$

In that case, the credal set $\mathscr{P} = \{P : P(A) \geq g(A), \forall A \subseteq S\}$ is characterized by the set-function g, that is, it can be described by assigning optimal buying prices to events (viewed as 0-1 gambles) only. Mind that not all credal sets can be characterized in this way. They generally require the assignment of buying (or selling) prices to general gambles. A sufficient condition for a monotonic set function to be Walley-coherent is the supermodularity condition: $g(A \cup B) + g(A \cap B) \geq g(A) + g(B)$. Such a function g is a called a *convex capacity*. So it is clear that other set-functions met in this chapter and the previous one are Walley-coherent as well, such as belief functions (equivalently plausibility functions) and necessity measures (equivalently possibility measures), which can represent specific credal sets.

Interestingly, Walley-coherence can be viewed as a generalization of deductive closure to families of weighted propositions. Let $\mathscr{K}$ be a consistent set of propositions $A_0, A_1, \ldots, A_k$, and suppose we assign the buying prices $pr(A_i) = 1, i = 0, \ldots k$, then $P_*(A) = 1$ if and only if $\mathscr{K} \models A$.

More about imprecise probability theories can be found in Walley [1991]'s book and their relevance for uncertainty management in artificial intelligence is discussed in [Walley, 1996], where the position of belief functions and possibility measures in the landscape is pointed out. More recent books on the topics are the collection of introductory papers edited by Augustin et al [2014], and the mathematically oriented monograph on lower previsions by de Cooman and Troffaes [2014].

3.2 Two Types of Conditioning

In the framework of imprecise probabilities, there are several ways of extending the Bayesian conditioning of probability theory. It reflects the fact that the two usual tasks performed by Bayes rule, that is prediction and revision, can no longer be performed by the same conditioning rule [Dubois and Prade, 1997b].

3.2.1 Prediction

When a credal set represents generic knowledge, Bayesian prediction or plausible inference is achieved by performing a form of sensitivity analysis on probabilistic conditioning, a rule proposed in [Walley, 1991; Fagin and Halpern, 1991]. Let $\mathscr{P}$ be a credal set on S. It induces lower and upper bounds $P_*(A)$ and $P^*(A)$ of the probability of each proposition A. In the presence of new pieces of information about

a singular case, summarized by the context C, the belief of the agent that proposition A holds for the case at hand is represented by the interval $[P_*(A \mid C), P^*(A \mid C)]$ defined by

$$P_*(A \mid C) = \inf\{P(A \mid C) \text{ s.t. } P(C) > 0, P \in \mathscr{P}\}$$
$$P^*(A \mid C) = \sup\{P(A \mid C) \text{ s.t. } P(C) > 0, P \in \mathscr{P}\}.$$

Note that it is possible that interval $[P_*(A \mid C), P^*(A \mid C)]$ is larger than $[P_*(A), P^*(A)]$, which means that there is a deficit of information given by the credal set $\mathscr{P}$ in the specific context C, while there is more in more general contexts. This is called the dilation effect [Seidenfeld and Wasserman, 1993]. It reflects the fact that in the presence of incomplete information, the more observations are available on a singular case, the less relevant to this case is generic information about the population of cases, because the less the new one can be viewed as representative of this population. In the case of Bayes rule applied to a known frequentist distribution, this dilation effect does not appear because in any case a single number is obtained. However, this value becomes all the more dubious as the number of cases similar to the one under study in the population justifying the frequentist distribution becomes smaller and smaller as we condition on a more specific context.

If $\mathscr{P}$ is the credal set associated to a convex capacity (hence, belief functions, necessity measures as well) the upper and lower conditional functions take the remarkable forms [Fagin and Halpern, 1991]:

$$P_*(A \mid C) = \frac{P_*(A \cap C)}{P_*(A \cap C) + P^*(\overline{A} \cap C)}; \quad P^*(A \mid C) = \frac{P^*(A \cap C)}{P^*(A \cap C) + P_*(\overline{A} \cap C)} \tag{37}$$

It is easy to see that $P^*(A \mid C) = 1 - P_*(\overline{A} \mid C)$, and these formula extend probabilistic conditioning, in the sense that $P_*(A \mid C)$ is a function of $P_*(A \cap C)$ and $P^*(\overline{C} \cup A)$ (and similarly for $P^*(A \mid C)$). It is clear that this form of conditioning does not correspond to the idea of enriching generic information by new observations, i.e., the latter do not alter the credal set. We just extract from it information that fits the available evidence, in the spirit of De Finetti.

In the theory of belief functions, the above form of conditioning can be justified in terms of their mass functions, positive weights $m(E)$ assigned to subsets E of S. When a mass function represents generic knowledge, $m(E)$ may be, e.g., the proportion of individuals for which imprecise proposition E holds, in the whole population. In this setting, prediction in context C consists in evaluating mass function $m(\cdot \mid C)$ induced by m in context C summarizing the available singular information. Three cases can be considered [de Campos et al, 1990]:

1. $E \subseteq C$: in that case, $m(E)$ remains committed to E;
2. $E \cap C = \emptyset$: in that case, $m(E)$ is no longer relevant and is discarded;
3. $E \cap C \neq \emptyset$ and $\overline{E} \cap C \neq \emptyset$: in that case, a part $\alpha_E \cdot m(E)$ of $m(E)$ remains committed to $E \cap C$ and the rest, i.e., $(1 - \alpha_E) \cdot m(E)$, is committed to $\overline{E} \cap C$. But the proportion α_E is unknown.

The third case corresponds to incomplete information E which neither confirms, nor contradicts C. We do not have information to determine if, in each of the situations

corresponding to these observations, C is true or not. Assume that one knows the proportions $\{\alpha_E, E \subseteq S\}$. We always have $\alpha_E = 1$ in the first case and $\alpha_E = 0$ in the second case. One thus constructs a mass function $m_\alpha(\cdot \mid C)$. We can remark that renormalization of the resulting mass function is necessary whenever $Pl(C) < 1$: each mass is then divided by $Pl(C)$. Denoting by $Bel_\alpha(A \mid C)$ and $Pl_\alpha(A \mid C)$ the belief and plausibility obtained by focalization on C with vector of proportions α, we can define the conditional degrees of belief and of plausibility given C as

$$Bel(A \mid C) = \inf_\alpha Bel_\alpha(A \mid C); \quad Pl(A \mid C) = \sup_\alpha Pl_\alpha(A \mid C). \tag{38}$$

These definitions yield the following special cases of Bayesian conditioning for imprecise probability (37):

$$Bel(A \mid C) = \inf\{P(A \mid C) \text{ s.t. } P(C) > 0, P \geq Bel\} = \frac{Bel(A \cap C)}{Bel(A \cap C) + Pl(\overline{A} \cap C)}; \tag{39}$$

$$Pl(A \mid C) = \sup\{P(A \mid C) \text{ s.t. } P(C) > 0, P \geq Bel\} = \frac{Pl(A \cap C)}{Pl(A \cap C) + Bel(\overline{A} \cap C)}. \tag{40}$$

We still obtain belief and plausibility functions[2] (see the non-trival proofs by Jaffray [1992] and Paris [1994]). Let us notice that if $Bel(C) = 0$ and $Pl(C) = 1$ (total ignorance about C) then all focal sets of m overlap C but C does not contain any of them. In that case, $Bel(A \mid C) = 0$ and $Pl(A \mid C) = 1, \forall A \neq S, \emptyset$: nothing can be inferred in context C.

3.2.2 Revision

In the framework of imprecise probabilities, a simple brute force approach to revision of a credal set $\mathscr{P}$ by an information item C consists in enforcing the additional constraint $P(C) = 1$ to $\mathscr{P}$, namely restrict the latter, and update the upper and lower probabilities of events accordingly:

$$P_*(A \mid\mid C) = \inf\{P(A \mid C) \text{ s.t. } P(C) = 1, P \in \mathscr{P}\}; \tag{41}$$

$$P^*(A \mid\mid C) = \sup\{P(A \mid C) \text{ s.t. } P(C) = 1, P \in \mathscr{P}\}. \tag{42}$$

Clearly, it is supposed, in contrast with the assumption in the prediction problem, that the new item of information is of the same nature as the original credal set, and can be modelled by the credal set $\{P : P(C) = 1\}$ (it can be frequentist or subjectivist).

However, by doing so, it may be that the intersection of the two credal sets, i.e., $\{P \in \mathscr{P} \text{ s.t. } P(C) = 1\}$ is empty. This is for instance most of the time the case in the standard probabilistic setting since $\mathscr{P}$ reduces to a singleton. The way out is to

[2] When applied to necessity and plausibility measures, these two formulas also preserve consonance and yield another form of conditional possibility and necessity [Dubois and Prade, 1997a].

apply the maximum likelihood principle [Gilboa and Schmeidler, 1992], selecting the most likely probability functions in $\mathscr{P}$, replacing condition $P(C) = 1$ by $P(C) = P^*(C)$ in the above definition of conditioning:

$$P_*(A \,||\, C) = \inf\{P(A \mid C) \text{ s.t. } P(C) = P^*(C), P \in \mathscr{P}\}; \tag{43}$$

$$P^*(A \,||\, C) = \sup\{P(A \mid C) \text{ s.t. } P(C) = P^*(C), P \in \mathscr{P}\}. \tag{44}$$

For convex capacities, it holds that $P^*(A \,||\, C) = \frac{P^*(A \cap C)}{P^*(C)}$, which generalizes Dempster rule of conditioning. In the belief function setting, this form of conditioning systematically assumes that $\alpha_E = 1$ whenever $E \cap C \neq \emptyset$ in $Bel_\alpha(A \mid C)$ and $Pl_\alpha(A \mid C)$. From the perspective of Shafer and Smets, mass function m does not represent generic information, but uncertain singular information, such as unreliable testimonies or inconclusive pieces of evidence about a specific situation. The existence of two forms of conditioning in the theory of belief functions can thus be explained by the difference between generic and singular information.

As a general setting for the numerical representation of uncertainty, liable of various interpretations, and encompassing other theories of uncertainty as formal particular cases, imprecise probabilities receive an increasing attention and foster a number of theoretical works (for instance, in de Cooman and Hermans [2008] bridges are built between Walley's approach to imprecise probabilities and the game-theoretic view of probability by Shafer and Vovk [2001]). Practical representation methods in artificial intelligence are also studied, for instance the imprecise probability version of Bayesian nets, including dedicated uncertainty propagation algorithms [Cozman, 2000; de Campos and Cozman, 2005; Cozman and Mauá, 2017].

4 Conclusion

Artificial Intelligence, when focusing on representation and reasoning with imperfect information, was naturally bound to realize that classical logic on the one hand, and precise probabilities on the other hand, were separately insufficient to deal with this issue. Alternative formal frameworks have emerged in the last 40 years or so to that effect, that this chapter partially accounts for. These frameworks are numerous and often complement each other rather than compete, even if research in this area remains fragmented. Nevertheless, these alternative theories of uncertain, incomplete or conflicting information offer a very rich range of formalisms. It is important to correctly understand their potentials and limitations prior to appropriately exploiting them. These frameworks can be qualitative (like possibilistic logic, dicussed in the previous chapter) or quantitative (like belief functions and imprecise probabilities). A significant effort is still needed before a full-fledged unification of the various approaches is achieved, and the links with neighboring disciplines like statistics are fully established, in order to master their use in applications.

References

Antoine V, Quost B, Masson MH, Denoeux T (2012) CECM: Constrained evidential c-means algorithm. Computational Statistics & Data Analysis 56(4):894–914

Antoine V, Quost B, Masson MH, Denoeux T (2014) CEVCLUS: evidential clustering with instance-level constraints for relational data. Soft Computing 18(7):1321–1335

Appriou A (1991) Probabilités et incertitude en fusion de données multi-senseurs. Revue Scientifique et Technique de la Défense 11:27–40

Appriou A (1998) Uncertain data aggregation in classification and tracking processes. In: Bouchon-Meunier B (ed) Aggregation and Fusion of imperfect information, Physica-Verlag, Heidelberg, pp 231–260

Augustin T, Coolen F, de Cooman G, Troffaes M (eds) (2014) Introduction to Imprecise Probabilities. Wiley

Baudrit C, Dubois D (2006) Practical representations of incomplete probabilistic knowledge. Computational Statistics & Data Analysis 51(1):86–108

Bauer M (1997) Approximation algorithms and decision making in the Dempster-Shafer theory of evidence – an empirical study. Int J Approx Reas 17:217–237

Berger JO (1994) An overview of robust Bayesian analysis. Test 3:5–124, with discussion

Bezdek J (1981) Pattern Recognition with fuzzy objective function algorithm. Plenum Press, New-York

Bi Y (2012) The impact of diversity on the accuracy of evidential classifier ensembles. Int J of Approximate Reasoning 53(4):584–607

Bi Y, Guan J, Bell D (2008) The combination of multiple classifiers using an evidential reasoning approach. Artificial Intelligence 172(15):1731–1751

Cattaneo MEGV (2011) Belief functions combination without the assumption of independence of the information sources. Int J of Approximate Reasoning 52(3):299–315

Couso I, Dubois D, Sanchez L (2014) Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables. SpringerBriefs in Computational Intelligence, Springer, http://www.springerlink.com

Cozman FG (2000) Credal networks. Artificial Intelligence 120:199–233

Cozman FG, Mauá DD (2017) On the complexity of propositional and relational credal networks. Int J Approx Reasoning 83:298–319

de Campos CP, Cozman FG (2005) The inferential complexity of Bayesian and credal networks. In: Kaelbling LP, Saffiotti A (eds) Proc. 19th Int. Joint Conf. on Artificial Intelligence (IJCAI'05), AAAI Press, pp 1313–1318

de Campos LM, Lamata MT, Moral S (1990) The concept of conditional fuzzy measure. Int J of Intell Syst 5:237–246

de Campos LM, Huete JF, Moral S (1994) Probability intervals: a tool for uncertain reasoning. Int J of Uncertainty, Fuzziness and Knowledge-Based Systems 2:167–196

de Cooman G, Hermans F (2008) Imprecise probability trees: Bridging two theories of imprecise probability. Artificial Intelligence 172:1400–1427

de Cooman G, Troffaes M (2014) Lower previsions. Wiley

Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. The Annals of Statistics 28:325–339

Dempster AP (1968) Upper and lower probabilities generated by a random closed interval. Annals of Mathematical Statistics 39(3):957–966

Denœux T (1995) A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans on Syst, Man and Cybern 25(05):804–813

Denœux T (1997) Analysis of evidence-theoretic decision rules for pattern classification. Pattern Recognition 30(7):1095–1107

Denœux T (1999) Reasoning with imprecise belief structures. Int J of Approximate Reasoning 20:79–111

Denœux T (2000a) Modeling vague beliefs using fuzzy-valued belief structures. Fuzzy Sets and Systems 116(2):167–199

Denœux T (2000b) A neural network classifier based on Dempster-Shafer theory. IEEE Trans on Syst, Man and Cybern A 30(2):131–150

Denœux T (2001) Inner and outer approximation of belief structures using a hierarchical clustering approach. Int J of Uncertainty, Fuzziness and Knowledge-Based Systems 9(4):437–460

Denœux T (2008) Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. Artificial Intelligence 172:234–264

Denœux T (2009) Extending stochastic ordering to belief functions on the real line. Information Sciences 179:1362–1376

Denœux T (2016) evclust: Evidential Clustering. URL `https://CRAN.R-project.org/package=evclust`, R package version 1.0.3

Denoeux T (2017) evclass: Evidential Distance-Based Classification. URL `https://CRAN.R-project.org/package=evclass`, R package version 1.1.1

Denœux T, Ben Yaghlane A (2002) Approximating the combination of belief functions using the fast Mœbius transform in a coarsened frame. Int J of Approximate Reasoning 31(1-2):77–101

Denoeux T, Kanjanatarakul O (2016) Beyond fuzzy, possibilistic and rough: An investigation of belief functions in clustering. In: Soft Methods for Data Science (Proc. SMPS 2016), Springer-Verlag, Berlin, Advances in Intelligent and Soft Computing, vol AISC 456, pp 157–164

Denœux T, Masson MH (2004) EVCLUS: Evidential clustering of proximity data. IEEE Trans on Syst, Man and Cybern B 34(1):95–109

Denœux T, Masson MH (2012) Evidential reasoning in large partially ordered sets. application to multi-label classification, ensemble clustering and preference aggregation. Annals of Operations Research 195(1):135–161

Denœux T, Smets P (2006) Classification using belief functions: the relationship between the case-based and model-based approaches. IEEE Transactions on Syst, Man and Cybern B 36(6):1395–1406

Denœux T, Zouhal LM (2001) Handling possibilistic labels in pattern classification using evidential reasoning. Fuzzy Sets and Systems 122(3):47–62

Denoeux T, Younes Z, Abdallah F (2010) Representing uncertainty on set-valued variables using belief functions. Artificial Intelligence 174(7-8):479–499

Denœux T, Kanjanatarakul O, Sriboonchitta S (2015) E*K*-NNclus: a clustering procedure based on the evidential *k*-nearest neighbor rule. Knowledge-based Systems 88:57–69

Denœux T, Sriboonchitta S, Kanjanatarakul O (2016) Evidential clustering of large dissimilarity data. Knowledge-based Systems 106:179–195

Destercke S, Burger T (2013) Toward an axiomatic definition of conflict between belief functions. IEEE Transactions on Cybernetics 43(2):585–596

Destercke S, Dubois D (2011) Idempotent conjunctive combination of belief functions: Extending the minimum rule of possibility theory. Information Sciences 181(18):3925–3945

Dubois D, Prade H (1982) A class of fuzzy measures based on triangular norms. a general framework for the combination of uncertain information. Int J General Syst 8(1):43–61

Dubois D, Prade H (1985) A note on measures of specificity for fuzzy sets. Int J General Syst 10(4):279–283

Dubois D, Prade H (1986) A set-theoretic view of belief functions: Logical operations and approximations by fuzzy sets. Int J General Syst 12:193–226

Dubois D, Prade H (1987) Properties of information measures in evidence and possibility theories. Fuzzy Sets and Systems 24:161–182

Dubois D, Prade H (1997a) Bayesian conditioning in possibility theory. Fuzzy Sets and Systems 92:223–240

Dubois D, Prade H (1997b) Focusing vs. belief revision: A fundamental distinction when dealing with generic knowledge. In: Gabbay DM, Kruse R, Nonnengart A, Ohlbach HJ (eds) Qualitative and Quantitative Practical Reasoning (Proc. ECSQARU-FAPR'97), Springer, LNCS, vol 1244, pp 96–107

Dubois D, Ramer A (1993) Extremal properties of belief measures in the theory of evidence. Int J of Uncertainty, Fuzziness and Knowledge-Based Systems 1(1):57–68

Dubois D, Prade H, Smets P (2001) "Not impossible" vs. "guaranteed possible" in fusion and revision. In: Benferhat S, Besnard P (eds) Symbolic and Quantitative Approaches to Reasoning with Uncertainty (Proc. ECSQARU'01), Springer, LNCS, vol 2143, pp 522–531

Dubois D, Prade H, Smets P (2008) A definition of subjective possibility. Int J Approx Reasoning 48:352–364

Fabre S, Appriou A, Briottet X (2001) Presentation and description of two classification methods using data fusion based on sensor management. Information Fusion 2(1):49–71

Fagin R, Halpern J (1991) A new approach to updating beliefs. In: Bonissone PP, Henrion M, Kanal LN, Lemmer JF (eds) Uncertainty in Artificial Intelligence, vol 6, North-Holland, Amsterdam, pp 347–374

Gilboa I, Schmeidler D (1989) Maxmin expected utility with a non-unique prior. J of Mathematical Economics 18:141–153

Gilboa I, Schmeidler D (1992) Updating ambiguous beliefs. In: Moses Y (ed) Proc. of the 4th Conf. on Theoretical Aspects of Reasoning about Knowledge (TARK'92), Morgan Kaufmann, pp 143–162

Giles R (1982) Foundations for a theory of possibility. In: Gupta MM, Sanchez E (eds) Fuzzy Information and Decision Processes, North-Holland, pp 183–195

Good IJ (1962) Subjective probability as the measure of a non-measurable set. In: Nagel E, Suppes P, Tarski A (eds) Handbook of the History of Logic, Stanford University Press, pp 319–329

Grabisch M (2009) Belief functions on lattices. Int J of Intell Syst 24:76–95

Harmanec D (1999) Faithful approximations of belief functions. In: Laskey KB, Prade H (eds) Uncertainty in Artificial Intelligence (Proc. UAI99), Stockholm

Harmanec D, Klir GJ (1994) Measuring total uncertainty in Dempster-Shafer theory: A novel approach. Int J General Syst 22(4):405–419

Huber P (1981) Robust statistics. Wiley, New York

Jaffray JY (1989) Linear utility theory for belief functions. Operations Research Letters 8(2):107 – 112

Jaffray JY (1992) Bayesian updating and belief functions. IEEE Trans on Systems, Man, and Cybernetics 22:1144–1152

Jiroušek R, Shenoy PP (2016) Entropy of belief functions in the dempster-shafer theory: A new perspective. In: Vejnarová J, Kratochvíl V (eds) Belief Functions: Theory and Applications (Proc. BELIEF 2016), Springer, Berlin, pp 3–13

Kennes R (1992) Computational aspects of the Möbius transformation of graphs. IEEE Trans on Systems, Man, and Cybernetics 22:201–223

Klir GJ, Wierman MJ (1999) Uncertainty-Based Information. Elements of Generalized Information Theory. Springer-Verlag, New-York

Lefèvre E, Colot O, Vannoorenberghe P (2002) Belief function combination and conflict management. Information Fusion 3(2):149–162

Lelandais B, Ruan S, Denœux T, Vera P, Gardin I (2014) Fusion of multi-tracer PET images for dose painting. Medical Image Analysis 18(7):1247–1259

Li F, Li S, Denœux T (2018) k-cevclus: Constrained evidential clustering of large dissimilarity data. Knowledge-Based Systems 142:29–44

Lian C, Ruan S, Denœux T (2015) An evidential classifier based on feature selection and two-step classification strategy. Pattern Recognition 48:2318–2327

Lian C, Ruan S, Denœux T (2016) Dissimilarity metric learning in the belief function framework. IEEE Transactions on Fuzzy Systems 24(6):1555–1564

Lingras P, Peters G (2012) Applying rough set concepts to clustering. In: Peters G, Lingras P, Ślezak D, Yao Y (eds) Rough Sets: Selected Methods and Applications in Management and Engineering, Springer-Verlag, London, UK, pp 23–37

Lowrance JD, Garvey TD, Strat TM (1986) A framework for evidential-reasoning systems. In: Proc. National AI Conference (AAAI'86), AAAI Press, vol 2, pp 896–903

Maeda Y, Ichihashi H (1993) An uncertainty measure with monotonicity under the random set inclusion. Int J General Syst 21(4):379–392

Martin A, Osswald C, Dezert J, Smarandache F (2008) General combination rules for qualitative and quantitative beliefs. J of Advances in Information Fusion 3(2):67–82

Masson MH, Denœux T (2008) ECM: an evidential version of the fuzzy c-means algorithm. Pattern Recognition 41(4):1384–1397

Masson MH, Denœux T (2009) RECM: relational evidential c-means algorithm. Pattern Recognition Letters 30:1015–1026

Mercier D, Quost B, Denœux T (2008) Refined modeling of sensor reliability in the belief function framework using contextual discounting. Information Fusion 9(2):246–258

Mercier D, Cron G, Denœux T, Masson MH (2009) Decision fusion for postal address recognition using belief functions. Expert Syst with Appl 36(3):5643–5653

Mercier D, Lefèvre E, Delmotte F (2012) Belief functions contextual discounting and canonical decompositions. Int J Approx Reas 53(2):146–158

Mercier D, Pichon F, Lefèvre E (2016) Corrigendum to "Belief functions contextual discounting and canonical decompositions" [Int. J. Approx. Reas. 53 (2012) 146–158]. Int J Approx Reas 70:137 – 139

Moral S, Wilson N (1994) Markov-chain Monte-Carlo algorithms for the calculation of Dempster-Shafer belief. In: Proc. of the Twelfth National Conference on Artificial intelligence (AAAI-94), vol 1, pp 269–274

Moral S, Wilson N (1996) Importance sampling Monte-Carlo algorithms for the calculation of Dempster-Shafer belief. In: Proc. of Int. Conf. on Inform. Proc. and Manag. of Uncertainty (Proc. IPMU'96), Granada, Spain, vol III, pp 1337–1344

von Neumann J, Morgenstern O (1944) Theory Games and Economic Behavior. Princeton University Press, Princeton, NJ

Nguyen H (2006) An Introduction to Random Sets. Chapman and Hall/CRC Press, Boca Raton, Florida

Nguyen HT (1978) On random sets and belief functions. J of Math Anal and Appl 65:531–542

Paris J (1994) The Uncertain Reasoner' Companion. Cambridge University Press

Pearl J (1990) Reasoning with belief functions: An analysis of compatibility. Int J Approx Reas 4(5):363 – 389

Petit-Renaud S, Denœux T (2004) Nonparametric regression analysis of uncertain and imprecise data using belief functions. Int J of Approximate Reasoning 35(1):1–28

Pichon F, Denœux T (2010) The unnormalized Dempster's rule of combination: a new justification from the least commitment principle and some extensions. J of Automated Reas 45(1):61–87

Pichon F, Denœux T, Dubois D (2012) Relevance and truthfulness in information correction and fusion. Int J of Approximate Reasoning 53(2):159–175

Pichon F, Mercier D, , Delmotte F (2016) Proposition and learning of some belief function contextual correction mechanisms. Int J Approx Reas 72:4 – 42

Quost B, Denœux T, Masson MH (2007) Pairwise classifier combination using belief functions. Pattern Recognition Letters 28(5):644–653

Quost B, Masson MH, Denœux T (2011) Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules. Int J Approx Reas 52(3):353–374

Ramer A (1987) Uniqueness of information measure in the theory of evidence. Fuzzy Sets and Systems 24:183–196

Ramer A, Klir GJ (1993) Measures of discord in the Dempster-Shafer theory. Information Sci 67:35–50
Seidenfeld T, Wasserman L (1993) Dilation for sets of probabilities. The Annals of Statistics 21:1139–1154
Shafer G (1973) Allocations of probability: A theory of partial belief. PhD thesis, Princeton University
Shafer G (1976) A Mathematical Theory of Evidence. Princeton Univ. Press
Shafer G (1979) Allocations of probability. Annals of Probability 7(5):827–839
Shafer G (1981) Constructive probability. Synthese 48(1):1–60
Shafer G (1987) Probability judgment in artificial intelligence and expert systems. Statistical Science 2(1):3–44
Shafer G (1990) Perspectives in the theory and practice of belief functions. Int J Approx Reas 4:323–362
Shafer G (2016a) Dempster's rule of combination. Int J Approx Reas 79:26–40
Shafer G (2016b) A Mathematical Theory of Evidence turns 40. Int J Approx Reas 79:7–25
Shafer G (2016c) The problem of dependent evidence. Int J Approx Reas 79:41–44
Shafer G, Vovk V (2001) Probability and Finance: It's Only a Game! Wiley, New York
Shapley LS (1953) A value for n-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the Theory of Games, volume II, Annals of Mathematical Studies series, vol 28, Princeton University Press, pp 307–317
Shenoy PP (1989) A valuation-based language for expert systems. Int J of Approximate Reasoning 3:383–411
Smets P (1981) The degree of belief in a fuzzy event. Information Sci 25:1–19
Smets P (1983) Information content of an evidence. Int J of Man-Machine Studies 19:33–43
Smets P (1990a) The combination of evidence in the Transferable Belief Model. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(5):447–458
Smets P (1990b) Constructing the pignistic probability function in a context of uncertainty. In: Henrion M, Shachter RD, Kanal LN, Lemmer J (eds) Uncertainty in Artificial Intelligence 5, Elsevier Science Publ., pp 29–39
Smets P (1993) Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. Int J of Approximate Reasoning 9:1–35
Smets P (1995) The canonical decomposition of a weighted belief. In: Proc. Int. Joint Conf. on Artificial Intelligence, Morgan Kaufman, San Mateo, Ca, pp 1896–1901
Smets P (2005) Belief functions on real numbers. Int J Approx Reasoning 40:181–223
Smets P (2007) Analyzing the combination of conflicting belief functions. Information Fusion 8(4):387–412
Smets P, Kennes R (1994) The transferable belief model. Artificial Intelligence 66:191–234
Smith CAB (1961) Consistency in statistical inference and decision. J of the Royal Statistical Society B-23:1–37

Strat TM (1984) Continuous belief functions for evidential reasoning. In: Brachman RJ (ed) Proc. National Conf. on Artificial Intelligence (AAAI'84), Austin, Aug. 6-10, pp 308–313

Strat TM (1990) Decision analysis using belief functions. Int J Approx Reas 4(5–6):391–417

Tessem B (1993) Approximations for efficient computation in the theory of evidence. Artificial Intelligence 61:315–329

Walley P (1991) Statistical Reasoning with Imprecise Probabilities. Chapman and Hall

Walley P (1996) Measures of uncertainty in expert systems. Artificial Intelligence 83:1–58

Walley P, Fine T (1982) Towards a frequentist theory of upper and lower probability. The Annals of Statistics 10:741–761

Xu L, Krzyzak A, Suen CY (1992) Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Transactions on Syst, Man and Cybern 22(3):418–435

Yager RR (1986) The entailment principle for Dempster-Shafer granules. Int J of Intell Syst 1:247–262

Yager RR (1992) Decision making under Dempster-Shafer uncertainties. Int J General Syst 20(3):233–245

Yager RR, Liu LP (eds) (2008) Classic Works of the Dempster-Shafer Theory of Belief Functions. Springer Verlag, Heidelberg

Yen J (1990) Generalizing the Dempster-Shafer theory to fuzzy sets. IEEE Transactions on Syst, Man and Cybern 20(3):559–569

Zadeh LA (1979) Fuzzy sets and information granularity. In: M M Gupta RKR, Yager RR (eds) Advances in Fuzzy Sets Theory and Applications, North-Holland, Amsterdam, pp 3–18

Zouhal LM, Denœux T (1998) An evidence-theoretic k-NN rule with parameter optimization. IEEE Trans on Syst, Man and Cybern C 28(2):263–271