Evidential reasoning in large partially ordered sets. Application to multi-label classification, ensemble clustering and preference aggregation

Thierry Denoeux, Marie-Hélène Masson

Noname manuscript No.
(will be inserted by the editor)

Evidential reasoning in large partially ordered sets

Application to multi-label classification, ensemble clustering and preference aggregation

Thierry Denœux and Marie-Hélène Masson

the date of receipt and acceptance should be inserted later

Abstract The Dempster-Shafer theory of belief functions has proved to be a powerful formalism for uncertain reasoning. However, belief functions on a finite frame of discernment Ω are usually defined in the power set 2^Ω, resulting in exponential complexity of the operations involved in this framework, such as combination rules. When Ω is linearly ordered, a usual trick is to work only with intervals, which drastically reduces the complexity of calculations. In this paper, we show that this trick can be extrapolated to frames endowed with an arbitrary lattice structure, not necessarily a linear order. This principle makes it possible to apply the Dempster-Shafer framework to very large frames such as the power set, the set of partitions, or the set of preorders of a finite set. Applications to multi-label classification, ensemble clustering and preference aggregation are demonstrated.

Keywords Belief Functions, Dempster-Shafer theory, Evidence Theory, Lattices, Lattice Intervals, Classification, Clustering, Learning, Preference Relation, Preorder.

1 Introduction

The theory of belief functions originates from the pioneering work of Dempster [4,5] and Shafer [26]. In the 1990's, the theory was further developed by Smets [29,32], who proposed a non probabilistic interpretation (referred to as the "Transferable Belief Model") and introduced several new tools for information fusion and decision making. Big steps towards the application of belief functions to real-world problems involving many variables have been made with the introduction of efficient algorithms for computing marginals in valuation-based systems [27,28].

Although there has been some work on belief functions on continuous frames (see, e.g., [19,31]), the theory of belief functions has been mainly applied in the discrete setting. In this case, all functions introduced in the theory as representations of evidence (including mass, belief, plausibility and commonality functions) are defined from the Boolean lattice

Thierry Denœux
Heudiasyc, Université de Technologie de Compiègne, CNRS, E-mail: tdenoeux@hds.utc.fr

Marie-Hélène Masson
Heudiasyc, Université de Picardie Jules Verne, CNRS, E-mail: mmasson@hds.utc.fr

$(2^\Omega, \subseteq)$ to the interval $[0,1]$. Consequently, all operations involved in the theory (such as the conversion of one form of evidence to another, or the combination of two items of evidence using Dempster's rule) have exponential complexity with respect to the cardinality K of the frame Ω, which makes it difficult to use the Dempster-Shafer formalism in very large frames [35].

When the frame Ω is linearly ordered, a usual trick is to constrain the focal elements (i.e., the subsets of Ω such that $m(A) > 0$) to be *intervals* (see, for instance, [9]). The complexity of manipulating and combining mass functions is then drastically reduced from 2^K to K^2. Most formula of belief function theory work for intervals, because the set of intervals equipped with the inclusion relation has a *lattice structure*. As shown recently in [16], belief functions can be defined in any lattice, not necessarily Boolean. In this paper, this trick will be extended to the case of frames endowed with a lattice structure, not necessarily a linear order. As it will be shown, a lattice of intervals can be constructed, in which belief functions can be defined. This approach makes it possible to define belief functions in very large frames (such as the power set of a finite set Ω, the set of partitions of a finite set, or the set of preorders of a finite set) with manageable complexity.

The rest of this paper is organized as follows. The necessary background on belief functions and on lattices will first be recalled in Sections 2 and 3, respectively. Our main idea will then be exposed in Section 4. It will then be applied to three data analysis problems involving the definition and manipulation of belief functions on

1. The powerset of a finite set (Section 5), with application to *multi-label classification*;
2. The set of partitions of a finite set (Section 6), with application to *ensemble clustering*;
3. The set of preorders of a finite set (Section 7), with application to *preference aggregation* from pairwise comparisons.

Finally, Section 8 will conclude this paper. Note that the applications to multi-label classification and ensemble clustering have been described separately in [11,37] and [20,21], respectively. The present paper presents the approach in much greater generality and provides a unified view of the two previous applications. To demonstrate the generality of the new introduced framework, a third application to preference aggregation is dealt with. A unified decision making procedure based on commonalities is also introduced.

2 Belief Functions: Basic Notions

Let Ω be a finite set. A *(normalized) mass function* on Ω is a function $m : 2^\Omega \to [0,1]$ such that $m(\emptyset) = 0$ and

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{1}$$

The subsets A of Ω such that $m(A) > 0$ are called the *focal elements* of m. The normalization property can be relaxed by dropping the condition $m(\emptyset) = 0$. Only normalized mass functions will be considered in this paper. A mass function m is often used to model beliefs held by an agent about a variable X taking a single but ill-known value ω_0 in Ω [32]. The quantity $m(A)$ is then interpreted as the measure of the belief that is committed *exactly* to the hypothesis $\omega_0 \in A$. Full certainty corresponds to the case where $m(\{\omega_k\}) = 1$ for some $\omega_k \in \Omega$, while total ignorance is modeled by the *vacuous* mass function verifying $m(\Omega) = 1$.

To each normalized mass function m can be associated a *belief function bel* defined as follows:

$$bel(A) = \sum_{B \subseteq A} m(B), \tag{2}$$

for all $A \subseteq \Omega$. It is clear that $bel(\emptyset) = 0$ and $bel(\Omega) = 1$. Each quantity $bel(A)$ is interpreted as the *total degree of justified belief* [32] in the proposition "The true value ω_0 of X belongs to A". Conversely, m can be recovered from bel as

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} bel(B), \tag{3}$$

for all $A \subseteq \Omega$, where $|\cdot|$ denotes cardinality. Function m is said to be the *Möbius transform* of bel. For any function f from 2^Ω to $[0,1]$ such that $f(\emptyset) = 0$ and $f(\Omega) = 1$, the following conditions are known to be equivalent [26]:

1. The Möbius transform m of f is positive and verifies $\sum_{A \subseteq \Omega} m(A) = 1$.
2. f is totally monotone, i.e., for any $k \geq 2$ and for any family $A_1, \ldots, A_k$ in 2^Ω,

$$f\left(\bigcup_{i=1}^{k} A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} f\left(\bigcap_{i \in I} A_i\right).$$

Hence, bel defined by (2) is totally monotone.

Other functions related to m are the *plausibility function*, defined as

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - bel(\overline{A}), \quad \forall A \subseteq \Omega \tag{4}$$

and the *commonality function* (or co-Möbius transform of bel) defined as

$$q(A) = \sum_{B \supseteq A} m(B), \quad \forall A \subseteq \Omega. \tag{5}$$

Obviously, $q(\emptyset) = 1$ and $q(\Omega) = m(\Omega)$. Function m can be recovered from q using the following relation:

$$m(A) = \sum_{B \supseteq A} (-1)^{|B \setminus A|} q(B). \tag{6}$$

Functions m, bel, pl and q are thus in one-to-one correspondence and can be regarded as different facets of the same information.

When the reliability of a source is doubtful, the mass provided by this source can be *discounted* using the following operation:

$$\begin{cases} {}^{\alpha}m(A) = (1-\alpha)m(A) & \forall A \neq \Omega, \\ {}^{\alpha}m(\Omega) = (1-\alpha)m(\Omega) + \alpha, \end{cases} \tag{7}$$

where $0 \leq \alpha \leq 1$ is the *discount rate*. This coefficient is related to our confidence in the reliability of the source of information [30]. It can be interpreted as the plausibility that the source is unreliable. When α is equal to 1, the vacuous mass function is obtained. When $\alpha = 0$, m remains unchanged.

Let us now assume that we receive two normalized mass functions m_1 and m_2 from two distinct sources of information assumed to be reliable. Then m_1 and m_2 can be combined using Dempster's rule of combination defined as follows:

$$(m_1 \oplus m_2)(A) = \begin{cases} \dfrac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \kappa} & \text{if } A \neq \emptyset \\ 0 & \text{if } A = \emptyset, \end{cases} \tag{8}$$

where $\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ is the *degree of conflict* between the two mass functions, assumed to be strictly smaller than one. This rule is commutative, associative, and admits

the vacuous mass function as neutral element. Let $q_1 \oplus q_2$ denote the commonality function corresponding to $m_1 \oplus m_2$. It can be computed from q_1 and q_2, the commonality functions associated to m_1 and m_2, as follows:

$$(q_1 \oplus q_2)(A) = \frac{q_1(A) \cdot q_2(A)}{1-\kappa}, \tag{9}$$

for all non empty subset A of Ω, and $(q_1 \oplus q_2)(\emptyset) = 1$.

The conjunctive sum has a dual disjunctive rule [30], obtained by substituting union for intersection in (8), and dropping the normalization constant, which is no longer needed:

$$(m_1 \textcircled{\cup} m_2)(A) = \sum_{B \cup C = A} m_1(B) m_2(C), \quad \forall A \subseteq \Omega. \tag{10}$$

Denoting by $bel_1 \textcircled{\cup} bel_2$ the belief function corresponding to $m_1 \textcircled{\cup} m_2$, it can be shown that

$$(bel_1 \textcircled{\cup} bel_2)(A) = bel_1(A) \cdot bel_2(A), \quad \forall A \subseteq \Omega, \tag{11}$$

which is the counterpart of (9).

Given two mass functions m_1 and m_2 on the same frame of discernment Ω, we say that m_1 is a *specialization* of m_2 (or, equivalently, that m_2 is a *generalization* of m_1), if m_1 can be obtained from m_2 by transferring masses $m_2(B)$ to subsets of B, for all focal elements B of m_2 [12]. It is then more informative, or more committed. Formally, this property can be expressed as follows:

$$m_1(A) = \sum_{B \subseteq \Omega} S(A,B) m_2(B), \quad \forall A \subseteq \Omega, \tag{12}$$

where $S : 2^\Omega \times 2^\Omega \rightarrow [0,1]$ verifies

$$\sum_{A \subseteq \Omega} S(A,B) = 1, \quad \forall B \subseteq \Omega,$$

and

$$S(A,B) > 0 \Rightarrow A \subseteq B, \quad A, B \subseteq \Omega.$$

3 Belief Functions in General Lattices

As shown by Grabisch [16], the theory of belief functions can be defined not only in Boolean lattices, but in any lattice, not necessarily Boolean. We will first recall some basic definitions about lattices. Grabisch's results used in this work will then be summarized.

3.1 Lattices

A review of lattice theory can be found in [22]. The following presentation follows [16].

Let L be a finite set and $\leq$ a partial ordering (i.e., a reflexive, antisymmetric and transitive relation) on L. The structure $(L, \leq)$ is called a *poset*. We say that $(L, \leq)$ is a *lattice* if, for every $x, y \in L$, there is a unique greatest lower bound (denoted $x \wedge y$) and a unique least upper bound (denoted $x \vee y$). Operations $\wedge$ and $\vee$ are called the *meet* and *join* operations, respectively. For finite lattices, the greatest element (denoted $\top$) and the least element (denoted $\bot$) always exist. A strict partial ordering $<$ is defined from $\leq$ as $x < y$ if $x \leq y$ and

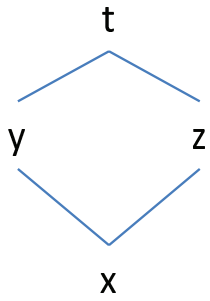

Fig. 1 Hasse diagram of a lattice on $\Omega = \{x,y,z,t\}$.

$x \neq y$. We say that x *covers* y if $y < x$ and there is no z such that $y < z < x$. An element x of L is an *atom* if it covers only one element and this element is $\bot$. It is a *co-atom* if it is covered by a single element and this element is $\top$.

Two lattices L and L' are *isomorphic* if there exists a bijective mapping f from L to L' such that $x \leq y \Leftrightarrow f(x) \leq f(y)$. For any poset $(L,\leq)$, we can define its dual $(L,\geq)$ by inverting the order relation. A lattice is *autodual* if it is isomorphic to its dual.

A lattice is *distributive* if $(x \vee y) \wedge z = (x \wedge z) \vee (y \wedge z)$ holds for all $x,y,z \in L$. For any $x \in L$, we say that x has a complement in L if there exists $x' \in L$ such that $x \wedge x' = \bot$ and $x \vee x' = \top$. L is said to be *complemented* if any element has a complement. Boolean lattices are distributive and complemented lattices. In a Boolean lattice, every element has a unique complement. Every Boolean lattice is isomorphic to $(2^{\Omega},\subseteq)$ for some set Ω. For the lattice $(2^{\Omega},\subseteq)$, we have $\wedge = \cap$, $\vee = \cup$, $\bot = \emptyset$ and $\top = \Omega$.

Example 1 Let $\Omega = \{x,y,z,t\}$ be a frame of discernment, and let $\leq$ be the following partial order: $x \leq y$, $x \leq z$, $y \leq t$, $z \leq t$. This order is represented by the Hasse diagram shown in Figure 1. In this representation, a line segments goes upward from a to b if b covers a. It is easy to see that $(\Omega,\leq)$ forms a lattice, isomorphic to the Boolean lattice $(2^{\Theta},\subseteq)$ with $|\Theta| = 2$. The least element is x, the greatest element is t, and y and z are both atoms and co-atoms. □

A *closure system* on a set Θ is a family $\mathscr{C}$ of subsets of Θ containing Θ, and closed under intersection. As shown in [22], any closure system $(\mathscr{C},\subseteq)$ is a lattice with $\wedge = \cap$ and $\vee = \sqcup$ defined by

$$A \sqcup B = \bigcap \{C \in \mathscr{C} | A \cup B \subseteq C\}, \quad \forall (A,B) \in \mathscr{C}^2. \tag{13}$$

3.2 Belief Functions on Lattices

Let $(L,\leq)$ be a finite poset having a least element, and let f be a function from L to $\mathbb{R}$. The *Möbius transform* of f is the function $m : L \to \mathbb{R}$ defined as the unique solution of the equation:

$$f(x) = \sum_{y \leq x} m(y), \quad \forall x \in L. \tag{14}$$

Function m can be expressed as:

$$m(x) = \sum_{y \leq x} \mu(y,x) f(y), \tag{15}$$

where $\mu(x,y) : L^2 \to \mathbb{R}$ is the *Möbius function*, which is defined inductively by:

$$\mu(x,y) = \begin{cases} 1 & \text{if } x = y, \\ -\sum\limits_{x \le t < y} \mu(x,t) & \text{if } x < y, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

The *co-Möbius transform* of f is defined as:

$$q(x) = \sum_{y \ge x} m(y), \tag{17}$$

and m can be recovered from q as:

$$m(x) = \sum_{y \ge x} \mu(x,y) q(y). \tag{18}$$

Let us now assume that $(L, \le)$ is a lattice. Following Grabisch [16], a function $bel : L \to [0,1]$ will be called a belief function on L if $bel(\bot) = 0$, $bel(\top) = 1$, and its Möbius transform is non negative.

As shown in [16], any belief function on $(L, \le)$ is totally monotone, i.e., for any $k \ge 2$ and for any family $x_1, \ldots, x_k$ in L,

$$bel\left(\bigvee_{i=1}^{k} x_i\right) \ge \sum_{\emptyset \ne I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} bel\left(\bigwedge_{i \in I} x_i\right).$$

However, the converse does not hold in general: a totally monotone function may not have a non negative Möbius transform.

As shown in [16], most results of Dempster-Shafer theory can be transposed in the general lattice setting. For instance, Dempster's rule can be extended by replacing $\cap$ by $\wedge$ in (8), and relation (9) between commonality functions is preserved. Similarly, we can extend the disjunctive rule (10) by substituting $\vee$ for $\cup$ in (10), and relation (11) still holds.

The discounting operation (7) and the notion of specialization (12) can be also generalized in straightforward ways. However, the extension of other notions from classical Dempster-Shafer theory may require additional assumptions on $(L, \le)$. For instance, the definition of the plausibility function *pl* as the dual of *bel* using (4) can only be extended to autodual lattices [16].

4 Belief functions with Lattice Intervals as Focal Elements

Let Ω be a finite frame of discernment. If the cardinality of Ω is very large, working in the Boolean lattice $(2^\Omega, \subseteq)$ may become intractable. This problem can be circumvented by defining the set of *propositions* as a strict subset of 2^Ω. As shown in Section 3, the Dempster-Shafer calculus can be applied in this restricted set of propositions as long as it has a lattice structure. To be meaningful, the definition of propositions should be based on some underlying structure of the frame of discernment.

When the frame Ω is linearly ordered, then a usual trick consists in assigning non zero masses only to intervals. Here, we propose to extend and formalize this approach, by considering the more general case where Ω has a lattice structure for some partial ordering $\le$.

The set of events is then defined as the set $\mathscr{I}_{\Omega,\leq}$ of lattice intervals in $(\Omega,\leq)$. We will show that $(\mathscr{I}_{\Omega,\leq},\subseteq)$ is then itself a lattice, in which the Dempster-Shafer calculus can be applied.

The lattice $(\mathscr{I}_{\Omega,\leq},\subseteq)$ of intervals of a lattice $(\Omega,\leq)$ will first be introduced more precisely in Section 4.1. The definition of belief functions in $(\mathscr{I}_{\Omega,\leq},\subseteq)$ will then be dealt with in Section 4.2.

4.1 The Lattice $(\mathscr{I}_{\Omega,\leq},\subseteq)$

Let Ω be a finite frame of discernment, and let $\leq$ be a partial ordering of Ω such that $(\Omega,\leq)$ is a lattice, with greatest element $\top$ and least element $\bot$. A subset I of Ω is a (lattice) interval if there exist elements a and b of Ω such that

$$I = \{x \in \Omega | a \leq x \leq b\}.$$

We then denote I as $[a,b]$. Obviously, Ω is the interval $[\bot,\top]$ and $\emptyset$ is the empty interval represented by $[a,b]$ for any a and b such that $a \leq b$ does not hold. Let $\mathscr{I}_{\Omega,\leq} \subseteq 2^{\Omega}$ be the set of intervals, including the empty set $\emptyset$:

$$\mathscr{I}_{\Omega,\leq} = \{[a,b] | a,b \in \Omega, a \leq b\} \cup \{\emptyset\}.$$

The intersection of two intervals is an interval:

$$[a,b] \cap [c,d] = \begin{cases} [a \vee c, b \wedge d] & \text{if } a \vee c \leq b \wedge d, \\ \emptyset & \text{otherwise.} \end{cases}$$

Consequently, $\mathscr{I}_{\Omega,\leq}$ is a closure system, and $(\mathscr{I}_{\Omega,\leq},\subseteq)$ is a lattice, with least element $\emptyset$ and greatest element Ω. We have

$$[a,b] \subseteq [c,d] \Leftrightarrow c \leq a \text{ and } b \leq d.$$

The meet operation is the intersection, and the join operation $\sqcup$ is defined by

$$[a,b] \sqcup [c,d] = [a \wedge c, b \vee d]. \tag{19}$$

Clearly, $[a,b] \subseteq [a,b] \sqcup [c,d]$ and $[c,d] \subseteq [a,b] \sqcup [c,d]$, hence $[a,b] \cup [c,d] \subseteq [a,b] \sqcup [c,d]$, but the inclusion is strict in general. We note that $(\mathscr{I}_{\Omega,\leq},\subseteq)$ is a subposet, but not a sublattice of $(2^{\Omega},\subseteq)$, because they do not share the same join operation.

The atoms of $(\mathscr{I}_{\Omega,\leq},\subseteq)$ are the singletons of Ω, while the co-atoms are intervals of the form $[\bot,x]$, where x is a co-atom of $(\Omega,\leq)$, or $[x,\top]$, where x is an atom of $(\Omega,\leq)$. The lattice $(\mathscr{I}_{\Omega,\leq},\subseteq)$ is usually neither autodual, nor Boolean.

Example 2 Let $(\Omega,\leq)$ be the lattice defined in Example 1. Figure 2 displays the corresponding lattice of intervals $(\mathscr{I}_{\Omega,\leq},\subseteq)$. There are ten distinct intervals of $(\Omega,\leq)$, including the empty set. The atoms of $(\mathscr{I}_{\Omega,\leq},\subseteq)$ are the singletons of Ω, and the co-atoms are $[x,y] = \{x,y\}$, $[x,z] = \{x,z\}$, $[y,t] = \{y,t\}$ and $[z,t] = \{z,t\}$. This lattice is complemented, but the complements are not unique: for instance, $\{x,z\} \cap \{y,t\} = \emptyset$ and $\{x,z\} \sqcup \{y,t\} = \Omega$, but we also have $\{x,z\} \cap \{t\} = \emptyset$ and $\{x,z\} \sqcup \{t\} = \Omega$: consequently, $\{y,t\}$ and $\{t\}$ are two complements of $\{x,z\}$. As a consequence, the lattice $(\mathscr{I}_{\Omega,\leq},\subseteq)$ is not Boolean. □

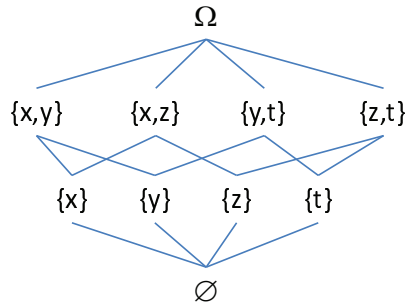

Fig. 2 The lattice $(\mathscr{I}_{\Omega,\leq},\subseteq)$ of intervals of the lattice $(\Omega,\leq)$ shown in Figure 1.

4.2 Belief Functions in $(\mathscr{I}_{\Omega,\leq},\subseteq)$

Let m be a mass function from $\mathscr{I}_{\Omega,\leq}$ to $[0,1]$. Belief and commonality functions can be defined in $(\mathscr{I}_{\Omega,\leq},\subseteq)$ as explained in Section 3. Conversely, m can be recovered from *bel* and q using (15) and (18), where the Möbius function μ depends on the lattice $(\mathscr{I}_{\Omega,\leq},\subseteq)$. As the cardinality of $\mathscr{I}_{\Omega,\leq}$ is at most proportional to K^2, where K is the cardinality of Ω, all these operations, as well as the conjunctive and disjunctive sums can be performed in polynomial time.

Example 3 Let us come back to Example 2. Given a mass function $m:\mathscr{I}_{\Omega,\leq}\rightarrow[0,1]$, the corresponding belief function can be computed as $bel(\emptyset)=0$, $bel(\Omega)=1$, and

$$bel(\{x\})=m(\{x\}),\quad bel(\{y\})=m(\{y\}),\quad bel(\{z\})=m(\{z\}),\quad bel(\{t\})=m(\{t\}),$$

$$bel(\{x,y\})=m(\{x\})+m(\{y\})+m(\{x,y\}),\quad bel(\{x,z\})=m(\{x\})+m(\{z\})+m(\{x,z\}),$$

$$bel(\{y,t\})=m(\{y\})+m(\{t\})+m(\{y,t\}),\quad bel(\{z,t\})=m(\{z\})+m(\{t\})+m(\{z,t\}).$$

By solving the above linear system, we easily find that m can be recovered from *bel* using (15), with the Möbius function defined for all A and B in $\mathscr{I}_{\Omega,\leq}$ by

$$\mu(A,B)=\begin{cases}(-1)^{|B\setminus A|} & \text{if } A\subseteq B\\ 0 & \text{otherwise.}\end{cases}$$

□

Given a mass function m on $(\mathscr{I}_{\Omega,\leq},\subseteq)$, we may define a function m^* on $(2^\Omega,\subseteq)$ as

$$m^*(A)=\begin{cases}m(A) & \text{if } A\in\mathscr{I}_{\Omega,\leq},\\ 0 & \text{otherwise.}\end{cases}$$

m^* will be called the *extension* of m in $(2^\Omega,\subseteq)$.

Let bel^* and q^* be the belief and commonality functions associated to m^*. It is obvious that $bel^*(I)=bel(I)$ and $q^*(I)=q(I)$ for all $I\in\mathscr{I}_{\Omega,\leq}$. Let m_1 and m_2 be two mass functions on $(\mathscr{I}_{\Omega,\leq},\subseteq)$, and let m_1^* and m_2^* be their extensions in $(2^\Omega,\subseteq)$. Because the meet operations

Table 1 Disjunctive combination of two mass functions m_1 and m_2 on the lattice $(\mathscr{I}_{\Omega,\leq},\subseteq)$ of Example 2.

A	$m_1(A)$	$m_2(A)$	$bel_1(A)$	$bel_2(A)$	$(bel_1 \textcircled{\cup} bel_2)(A)$	$(m_1 \textcircled{\cup} m_2)(A)$
$\emptyset$	0	0	0	0	0	0
$\{x\}$	0.1	0	0.1	0	0	0
$\{y\}$	0.2	0.3	0.2	0.3	0.06	0.06
$\{x,y\}$	0.3	0	0.6	0.3	0.18	0.12
$\{z\}$	0	0	0	0	0	0
$\{x,z\}$	0.2	0.4	0.3	0.4	0.12	0.12
$\{t\}$	0	0.1	0	0.1	0	0
$\{y,t\}$	0	0.1	0.2	0.5	0.1	0.04
$\{z,t\}$	0	0	0	0.1	0	0
Ω	0.2	0.1	1	1	1	0.66

are identical in $(\mathscr{I}_{\Omega,\leq},\subseteq)$ and $(2^\Omega,\subseteq)$, computing the conjunctive sum in any of these two lattices yields the same result, as we have

$$(m_1^* \oplus m_2^*)(A) = \begin{cases} (m_1 \oplus m_2)(A) & \text{if } A \in \mathscr{I}_{\Omega,\leq}, \\ 0 & \text{otherwise.} \end{cases}$$

However, computing the disjunctive sum in $(2^\Omega,\subseteq)$ or $(\mathscr{I}_{\Omega,\leq},\subseteq)$ is not equivalent, because the join operation in $(\mathscr{I}_{\Omega,\leq},\subseteq)$, defined by (19), is not identical to the union operation in 2^Ω. Consequently, when computing the disjunctive sum of m_1^* and m_2^*, the product $m_1^*(A)m_2^*(B)$ is transferred to $A \cup B$, whereas the product $m_1(A)m_2(B)$ is transferred to $A \sqcup B$ when combining m_1 and m_2. Let $(m_1 \textcircled{\cup} m_2)^*$ be the extension of $m_1 \textcircled{\cup} m_2$ in $(2^\Omega,\subseteq)$. As $A \sqcup B \supseteq A \cup B$, $(m_1 \textcircled{\cup} m_2)^*$ is thus a generalization (i.e., an *outer approximation* [12,8]) of $m_1^* \textcircled{\cup} m_2^*$. When masses are assigned to intervals of the lattice $(\Omega,\leq)$, doing the calculations in $(\mathscr{I}_{\Omega,\leq},\subseteq)$ can thus be seen an approximation of the calculations in $(2^\Omega,\subseteq)$, with a loss of information only when a disjunctive combination is performed.

Example 4 Table 1 shows two mass functions m_1 and m_2 on the lattice $(\mathscr{I}_{\Omega,\leq},\subseteq)$ defined in Example 2, as well as the corresponding belief functions, and the result $m_1 \textcircled{\cup} m_2$ of their disjunctive combination. Table 2 shows the extensions m_1^* and m_2^* of m_1 and m_2 in 2^Ω, as well as their disjunctive combination $m_1^* \textcircled{\cup} m_2^*$. It can be verified that $m_1^* \textcircled{\cup} m_2^*$ is a specialization of $(m_1 \textcircled{\cup} m_2)^*$, i.e., it is strictly more committed. Computing the disjunctive combination in the lattice $(\mathscr{I}_{\Omega,\leq},\subseteq)$ has thus resulted in a loss of information. □

To conclude this section, we may also remark that the reduction of complexity obtained by expressing beliefs in the lattice of intervals comes with a loss of expressive power. It is clear that, whereas any mass function m in $\mathscr{I}_{\Omega,\leq}$ has an extension m^* in 2^Ω, any mass function in 2^Ω with at least one focal element $A \subseteq \Omega$ not belonging to $\mathscr{I}_{\Omega,\leq}$ cannot be expressed in that lattice. Such a mass function could be approximated by transferring each mass $m(A)$ for $A \notin \mathscr{I}_{\Omega,\leq}$ to the smallest interval containing A. However, in the worst case, all information can be lost in this process.

For instance, considering again the lattice defined in Example 2, let m be the mass function on $(2^\Omega,\subseteq)$ such that $m(\{y,z,t\}) = 1$. This mass function expresses the opinion that the true value ω_0 in Ω is certainly not equal to x. It can be approximated in $\mathscr{I}_{\Omega,\leq}$ by transferring the unit mass of belief to the smallest interval containing $\{y,z,t\}$, which is Ω itself, resulting in the vacuous mass function. However, this is a somewhat extreme case. Complex real-world problems in which useful information can be expressed in a lattice of intervals will be described below.

Table 2 Disjunctive combination of two mass functions m_1^* and m_2^* on the Boolean lattice $(2^\Omega, \subseteq)$ with $\Omega = \{x,y,z,t\}$. Mass functions m_1^* and m_2^* are the extensions of m_1 and m_2 in Table 1. The elements of the lattice $\mathscr{I}_{\Omega,\leq}$ are preceded by an asterisk.

	A	$m_1^*(A)$	$m_2^*(A)$	$bel_1^*(A)$	$bel_2^*(A)$	$(bel_1^* \textcircled{\cup} bel_2^*)(A)$	$(m_1^* \textcircled{\cup} m_2^*)(A)$
*	$\emptyset$	0	0	0	0	0	0
*	$\{x\}$	0.1	0	0.1	0	0	0
*	$\{y\}$	0.2	0.3	0.2	0.3	0.06	0.06
*	$\{x,y\}$	0.3	0	0.6	0.3	0.18	0.12
*	$\{z\}$	0	0	0	0	0	0
*	$\{x,z\}$	0.2	0.4	0.3	0.4	0.12	0.12
	$\{y,z\}$	0	0	0.2	0.3	0.06	0
	$\{x,y,z\}$	0	0	0.8	0.7	0.56	0.26
*	$\{t\}$	0	0.1	0	0.1	0	0
	$\{x,t\}$	0	0	0.1	0.1	0.01	0.01
*	$\{y,t\}$	0	0.1	0.2	0.5	0.1	0.04
	$\{x,y,t\}$	0	0	0.6	0.5	0.3	0.07
*	$\{z,t\}$	0	0	0	0.1	0	0
	$\{x,z,t\}$	0	0	0.3	0.5	0.15	0.02
	$\{y,z,t\}$	0	0	0.2	0.5	0.1	0
*	Ω	0.2	0.1	1	1	1	0.3

4.3 Decision making

When working with belief functions in a Boolean Lattice $(2^\Omega, \subseteq)$, a usual decision rule is to select the singleton $\{\omega\}$ of Ω with the largest plausibility or, equivalently, with the largest commonality [7,3]. In the lattice $(\mathscr{I}_{\Omega,\prec}, \subseteq)$, the plausibility function is not defined, but the commonality function exists and its maximum can sometimes be computed efficiently without enumerating the elements of Ω, as will be shown below. A possible rule for decision making is thus to select the element of Ω with the largest commonality.

5 Reasoning with Set-valued Variables

In this section, we present a first application of the above scheme to the representation of knowledge regarding set-valued variables [11]. The general framework will be presented in Section 5.1, and it will be applied to multi-label classification in Section 5.2.

5.1 Evidence on Set-valued Variables

Let Θ be a finite set, and let X be a variable taking values in the power set 2^Θ. Such a variable is said to be set-valued, or *conjunctive* [12,36]. For instance, in diagnosis problems, Θ may denote the set of faults that can possibly occur in a system, and X the set of faults actually occurring at a given time, under the assumption that multiple faults can occur. In text classification, Θ may be a set of topics, and X the list of topics dealt with in a given text, etc.

Defining belief functions on the lattice $(2^{2^\Theta}, \subseteq)$ is practically intractable, because of the double exponential complexity involved. However, we may exploit the lattice structure induced by the ordering $\subseteq$ in $\Omega = 2^\Theta$, using the general approach outlined in Section 4 [11].

For any two subsets A and B of Θ such that $A \subseteq B$, the interval $[A,B]$ is defined as

$$[A,B] = \{C \subseteq \Theta | A \subseteq C \subseteq B\}.$$

The set of intervals of the lattice $(\Omega, \subseteq)$ is thus

$$\mathscr{I}_{\Omega,\subseteq} = \{[A,B] | A,B \in \Omega, A \subseteq B\} \cup \emptyset_\Omega,$$

where $\emptyset_\Omega$ denotes the empty sets of Ω (as opposed to the empty set of Θ). Clearly, $\mathscr{I}_{\Omega,\subseteq} \subseteq 2^\Omega = 2^{2^\Theta}$. The interval $[A,B]$ can be seen as the specification of an unknown subset C of Θ that *surely* contains all elements of A, and *possibly* contains elements of B. Alternatively, C surely contains *no* element of $\overline{B}$.

As the meet and join in the lattice $(\Theta, \subseteq)$ are set intersection and union, respectively, the corresponding operations in $(\mathscr{I}_{\Omega,\subseteq}, \subseteq)$ are

$$[A,B] \cap [C,D] = \begin{cases} [A \cup C, B \cap D] & \text{if } A \cup C \subseteq B \cap D, \\ \emptyset_\Omega & \text{otherwise} \end{cases}$$

and

$$[A,B] \sqcup [C,D] = [A \cap C, B \cup D].$$

As noticed in [17], any interval $[A,B]$ of subsets of $\Theta = \{\theta_1, \ldots, \theta_K\}$ can be represented by a vector $(u_1, \ldots, u_K) \in \{-1,0,1\}^K$, with

$$u_k = \begin{cases} 1 & \text{if } \theta_k \in A, \\ -1 & \text{if } \theta_k \in \overline{B}, \\ 0 & \text{otherwise.} \end{cases}$$

This encoding makes it possible to implement the $\cap$ and $\sqcup$ operations in a simple way using generalized truth tables. It also makes it clear that the cardinality of $\mathscr{I}_{\Omega,\subseteq}$ is equal to $3^K + 1$, which is much less than the 2^{2^K} elements of 2^Ω.

Example 5 Let $\Theta = \{a,b,c,d\}$ be the set of possible faults of a given system. Assume that several faults can occur simultaneously, and we receive two independent pieces of evidence:

- Item of evidence 1: fault a is surely present and faults $\{b,c\}$ may also be present, with confidence 0.7. This is represented by the following mass function:

$$m_1([\{a\}, \{a,b,c\}]) = 0.7, \quad m_1([\emptyset_\Theta, \Theta]) = 0.3.$$

- Item of evidence 2: fault c is surely present, and either faults $\{a,b\}$ (with confidence 0.8) or faults $\{a,d\}$ (with confidence 0.2) may also be present. This is represented by

$$m_2([\{c\}, \{a,b,c\}]) = 0.8, \quad m_2([\{c\}, \{a,c,d\}]) = 0.2.$$

The combination of m_1 and m_2 by Dempster's rule can be computed using the following table:

	$[\{a\}, \{a,b,c\}]$ 0.7	$[\emptyset_\Theta, \Theta]$ 0.3
$[\{c\}, \{a,b,c\}]$ 0.8	$[\{a,c\}, \{a,b,c\}]$ 0.56	$[\{c\}, \{a,b,c\}]$ 0.24
$[\{c\}, \{a,c,d\}]$ 0.2	$[\{a,c\}, \{a,c\}]$ 0.14	$[\{c\}, \{a,c,d\}]$ 0.06

Let $m_{12} = m_1 \oplus m_2$. We thus get

$$m_{12}([\{a,c\},\{a,b,c\}]) = 0.56, \quad m_{12}([\{c\},\{a,b,c\}]) = 0.24,$$

$$m_{12}([\{a,c\},\{a,c\}]) = 0.14, \quad m_{12}([\{c\},\{a,c,d\}]) = 0.06.$$

Based on this evidence, we can compute our degrees of belief in the following propositions:

– Fault a is present:

$$bel_{12}([\{a\},\Theta]) = 0.56 + 0.14 = 0.7$$

– Fault d is not present:

$$bel_{12}([\emptyset_\Theta, \overline{\{d\}}]) = bel_{12}([\emptyset_\Theta,\{a,b,c\}]) = 0.56 + 0.14 + 0.24 = 0.94$$

– Only faults a and c are present:

$$bel_{12}([\{a,c\},\{a,c\}]) = 0.14.$$

□

5.2 Application to Multi-label Classification

In [11,37], the above framework was applied to *multi-label classification* [39,34]. In this learning task, each object may belong simultaneously to several classes, contrary to standard single-label problems where objects belong to only one class. For instance, in image retrieval, each image may belong to several semantic classes such as "beach" or "urban". In such problems, the goal is to predict the value of the class variable for a new instance, based on a training set. As the class variable is set-valued, the framework developed in the previous section can be applied.

In order to construct a multi-label classifier, we generally assume the existence of a labeled training set, composed of n examples $(\mathbf{x}_i, Y_i)$, where $\mathbf{x}_i$ is a feature vector describing instance i, and Y_i is a label set for that instance, defined as a subset of the set Θ of classes [39, 34]. In practice, however, gathering such high quality information is not always feasible at a reasonable cost. In many problems, there is no ground truth for assigning unambiguously a label set to each instance, and the opinions of one or several experts have to be elicited. Typically, an expert will sometimes express lack of confidence for assigning exactly one label set.

The formalism described in Section 5.1 can easily be used to handle such situations. In the most general setting, the opinions of one or several experts regarding the set of classes that pertain to a particular instance i may be modeled by a mass function m_i in $(\mathscr{I}_{\Omega,\subseteq}, \subseteq)$. A less general, but arguably more operational option is to restrict m_i to be categorical, i.e., to have a single focal element $[A_i, B_i]$, with $A_i \subseteq B_i \subseteq \Theta$. The set A_i is then the set of classes that *certainly apply* to example i, while B_i is the set of classes that *possibly* apply to that instance. In a multiple expert setting, A_i might represent the set of classes indicated by *all* experts as relevant to describe instance i, while B_i would be the set of classes mentioned by *some* experts. The usual situation of precise labeling is recovered in the special case where $A_i = B_i$.

For instance, assume that instances are songs and classes are emotions generated by these songs. Upon hearing a song, an expert may decide that it certainly evokes happiness and certainly does not evoke sadness, but may be undecided regarding the other emotions

(such as quietness, anger, surprise, etc.). In that case, the song cannot be assigned a single label set, but we can associate to it the set of all label sets containing "happiness" and not containing "sadness", which has the form suggested above.

The evidential k nearest neighbor rule introduced in [6] can be extended to the multi-label framework as follows. Let $\Phi_k(\mathbf{x})$ denote the set of k nearest neighbors of a new instance described by feature vector $\mathbf{x}$, according to some distance measure d, and $\mathbf{x}_i$ an element of that set with label $[A_i, B_i]$. This item of evidence can be described by the following mass function in $(\mathscr{I}_{\Omega,\subseteq}, \subseteq)$:

$$\begin{aligned} m_i([A_i, B_i]) &= \beta \exp\left(-\gamma d(\mathbf{x}, \mathbf{x}_i)\right) = \alpha_i, \\ m_i([\emptyset_\Theta, \Theta]) &= 1 - \beta \exp\left(-\gamma d(\mathbf{x}, \mathbf{x}_i)\right) = 1 - \alpha_i, \end{aligned}$$

where β and γ are two parameters such that $0 < \beta < 1$ and $\gamma > 0$. These k mass functions are then combined into a single one m using Dempster's rule:

$$m = \oplus_{i=1}^{k} m_i. \tag{20}$$

For decision making, it was proposed in [11,37] to use the following rule. Let $\widehat{Y}$ be the predicted label set for instance $\mathbf{x}$. To decide whether to include each class $\theta \in \Theta$ or not, two quantities were computed: the degree of belief $bel([\{\theta\}, \Theta])$ that the true label set Y contains θ, and the degree of belief $bel([\emptyset, \overline{\{\theta\}}])$ that it does not contain θ. The set of predicted labels $\widehat{Y}$ was then defined as

$$\widehat{Y} = \{\theta \in \Theta \mid bel([\{\theta\}, \Theta]) \geq bel([\emptyset, \overline{\{\theta\}}])\}.$$

This method was shown in [11,38] to yield good performances compared to standard multi-label classification methods, especially when class labels are uncertain.

As noted in Section 4.3, an alternative way of making a decision is to find the set of labels with the greatest commonality. This approach may be formalized as follows. The commonality function corresponding to the combination (20) is given by:

$$q \propto \prod_{i=1}^{k} q_i, \tag{21}$$

where q_i is the commonality function associated to m_i. These individual commonalities can be simply expressed for any subset Y of Θ by:

$$q_i(Y) = \begin{cases} 1 & \text{if } A_i \subseteq Y \subseteq B_i, \\ 1 - \alpha_i & \text{otherwise.} \end{cases} \tag{22}$$

We thus have :

$$q(Y) \propto \prod_{i=1}^{k} (1 - \alpha_i)^{1-\delta_i}, \tag{23}$$

with

$$\delta_i = \begin{cases} 1 \text{ if } A_i \subseteq Y \subseteq B_i \\ 0 \text{ otherwise.} \end{cases}$$

For each focal element $[A_i, B_i]$, let us introduce the following notations:

$$a_{ij} = \begin{cases} 1 \text{ if } \theta_j \in A_i \\ 0 \text{ otherwise.} \end{cases}$$

and

$$b_{ij} = \begin{cases} 1 \text{ if } \theta_j \in B_i \\ 0 \text{ otherwise.} \end{cases}$$

In the same way, a subset Y of Θ will be represented by a K-dimensional vector $\mathbf{y}$ whose component are defined by $y_j = 1$ if $\theta_j \in Y$ and 0 otherwise. With these notations, the inclusion constraint $A_i \subseteq Y$ may be translated by:

$$\sum_{j=1}^{K} a_{ij} y_j = \sum_{j=1}^{K} a_{ij}.$$

Similarly, the constraint $Y \subseteq B_i$, or, equivalently, $\overline{B_i} \subseteq \overline{Y}$, may be written as:

$$\sum_{j=1}^{K} (1-b_{ij})(1-y_j) = \sum_{j=1}^{K} (1-b_{ij}).$$

Maximizing $q(Y)$ is equivalent to maximizing its logarithm, which is equal to:

$$\ln q(Y) = \sum_{i=1}^{k} (1-\delta_i) \ln(1-\alpha_i) + \text{constant}.$$

To find the set Y of greatest commonality, we can thus solve the following binary integer programming problem:

$$\min_{\mathbf{y} \in \{0,1\}^K, \delta \in \{0,1\}^k} \sum_{i=1}^{k} \delta_i \ln(1-\alpha_i), \tag{24}$$

subject to the constraints:

$$\begin{cases} \sum_{j=1}^{K} a_{ij} y_j \geq \delta_i \sum_{j=1}^{K} a_{ij} \quad \forall i = 1,k, \\ \sum_{j=1}^{K} (1-b_{ij})(1-y_j) \geq \delta_i \sum_{j=1}^{K} (1-b_{ij}) \quad \forall i = 1,k. \end{cases} \tag{25}$$

Note that the way in which the constraints (25) are written ensures that if δ_i is set to 1, Y is enforced to belong to $[A_i, B_i]$, and if δ_i is set to 0, there is no constraint on Y with respect to A_i and B_i. This method is more general and computationally more efficient than the method described in [11,38]; it has been found experimentally to yield similar results when applied to multi-label classification problem.

6 Belief Functions on Partitions

Ensemble clustering methods [18,14] aim at combining multiple clustering solutions or partitions into a single one, offering a better description of the data. In this section, we explain how to address this fusion problem using the general framework introduced in Section 4. Each clustering algorithm (or "clusterer") can be considered as a partially reliable source, giving an opinion about the true, unknown, partition of the objects. This opinion provides evidence in favor of a set of possible partitions. Moreover, we suppose that the reliability of each source is described by a confidence degree, either assessed by an external agent or evaluated using a class validity index. Manipulating beliefs defined on sets of partitions is intractable in the usual case where the number of potential partitions is high (for example, a

set composed of 6 elements has 203 potential partitions!) but it can be manageable using the lattice structure of partitions, as will be explained below. Note that, due to space limitations, only the main principles will be introduced. More details may be found in [20,21].

First, basic notions about the lattice of partitions of a set are recalled in Section 6.1, then our approach is explained and illustrated in Section 6.2 using a synthetic data set.

6.1 Lattice of Partitions

Let E denote a finite set of n objects. A partition p is a set of non empty, pairwise disjoint subsets $E_1,\ldots,E_k$ of E, such that their union is equal to E. Every partition p can be associated to an equivalence relation (i.e., a reflexive, symmetric, and transitive binary relation) on E, denoted by R_p, and characterized, for all $(a_i,a_j)\in E^2$, by:

$$R_p(a_i,a_j)=\begin{cases}1 & \text{if } a_i \text{ and } a_j \text{ belong to the same cluster in } p,\\ 0 & \text{otherwise.}\end{cases}$$

The set of all partitions of E, denoted Ω, can be partially ordered using the following ordering relation: partition p is said to be *finer* than partition p' on the same set E (denoted $p\preceq p'$) if the clusters of p can be obtained by splitting those of p' (or equivalently, if each cluster of p' is the union of some clusters of p). This partial ordering can be alternatively defined using the equivalence relations associated to p and p':

$$p\preceq p' \Leftrightarrow R_p(a_i,a_j)\leq R_{p'}(a_i,a_j),\quad \forall (a_i,a_j)\in E^2.$$

The set Ω endowed with the $\preceq$-order has a lattice structure [22]. In this lattice, the meet $p\wedge p'$ of two partitions p and p' is defined as the coarsest partition among all partitions finer than p and p'. The clusters of the meet $p\wedge p'$ are obtained by considering pairwise intersections between clusters of p and p'. The equivalence relation $R_{p\wedge p'}$ is simply obtained as the minimum of R_p and $R_{p'}$. The join $p\vee p'$ is similarly defined as the finest partition among those that are coarser than p and p'. The equivalence relation $R_{p\vee p'}$ is the *transitive closure* of the maximum of R_p and $R_{p'}$. The least element $\bot$ of that lattice is the *finest* partition, denoted $p_0=(1/2/\cdots/n)$, in which each object is a cluster. The greatest element $\top$ of $(\Omega,\preceq)$ is the *coarsest* partition denoted $p_E=(123\cdots n)$, composed of a single cluster containing the n objects. In this order, each partition precedes every partition derived from it by aggregating two of its clusters. Similarly, each partition covers all partitions derived by subdividing one of its clusters in two clusters.

Figures 3 and 4 show examples of partition lattices in the case where E is composed of three and four objects, respectively. The *atoms* of $(\Omega,\preceq)$ are the partitions preceded by p_0. There are $n(n-1)/2$ such partitions of $n-1$ clusters. Atoms are associated to matrices R_p with only one off-diagonal entry equal to 1. The co-atoms are dichotomies, i.e., partitions composed of two clusters.

For any partitions $\underline{p}$ and $\overline{p}$ of Ω such that $\underline{p}\preceq\overline{p}$, the interval $[\underline{p},\overline{p}]$ is defined as:

$$[\underline{p},\overline{p}]=\{p\in\Omega\,|\,\underline{p}\preceq p\preceq\overline{p}\}. \tag{26}$$

It is a particular set of partitions, namely, the set of all partitions finer than $\overline{p}$ and coarser than $\underline{p}$. The set of intervals of the lattice $(\Omega,\preceq)$ is:

$$\mathscr{I}_{\Omega,\preceq}=\{[\underline{p},\overline{p}]|\underline{p},\overline{p}\in\Omega,\underline{p}\preceq\overline{p}\}\cup\emptyset_\Omega.$$

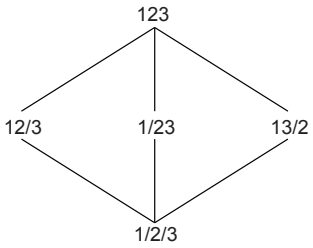

Fig. 3 Lattice of partitions of a three-element set.

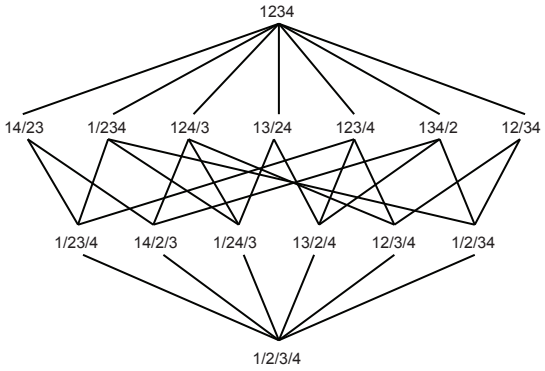

Fig. 4 Lattice of partitions of a four-element set.

In the lattice $(\mathscr{I}_{\Omega,\preceq},\subseteq)$, the least element $\perp$ is $\emptyset_\Omega$ and the greatest element $\top$ is Ω. The atoms are the singletons of Ω. The co-atoms are of the form $[p_0,p]$ with p a co-atom of $(\Omega,\preceq)$, or $[p,p_E]$ with p an atom of $(\Omega,\preceq)$. An example of such a lattice, in the case where E is composed of three objects, is shown in Figure 5.

Several forms of imprecise knowledge about a partition can be expressed in $(\mathscr{I}_{\Omega,\preceq},\subseteq)$. For instance, the intervals $[p_0,p]$ and $[p,p_E]$ represent the set of partitions finer and coarser, respectively, than a given partition p. Suppose now that the elements of a set $A\subseteq E$ are known to belong to the same cluster. This information can be represented by the interval $[p_A,p_E]$, where p_A denotes the partition in which the only elements that are clustered together are the elements of A:

$$p_A=\{A\}\cup\left\{\{a_i\}/a_i\in\bar{A}\right\}.$$

6.2 Application to Ensemble Clustering

In [21], the above approach was applied to ensemble clustering. The basic strategy can be summarized as follows:

1) Mass generation: Given r clusterers, build a collection of r mass functions m^1, m^2,..., m^r on the lattice of intervals $(\mathscr{I}_{\Omega,\preceq},\subseteq)$; the way of choosing the focal elements and allocating the masses from the results of several clusterers depends mainly on the applicative context and on the nature of the clusterers in the ensemble.

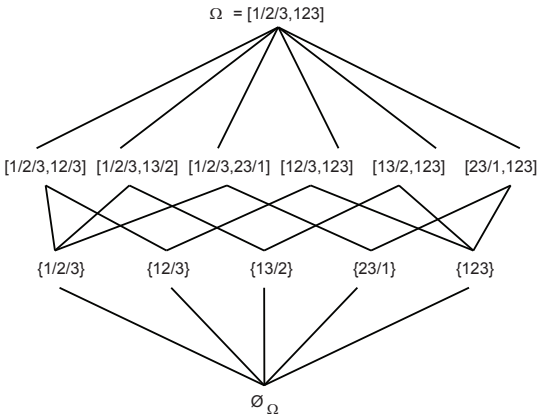

Fig. 5 Lattice of intervals of the lattice shown in Figure 3.

2) Aggregation: Combine the r mass functions into a single one using Dempster's rule. The result of this combination is a mass function m with focal elements $[\underline{p}_k, \overline{p}_k]$ and associated masses m_k, $k = 1, \ldots, s$. The equivalence relations corresponding to $\underline{p}_k$ and $\overline{p}_k$ will be denoted $\underline{R}_k$ and $\overline{R}_k$, respectively.
3) Decision making: as for multi-label classification, it is possible to solve a binary integer optimization problem for picking the best partition p with the greatest commonality. However, although theoretically possible, this approach necessitates a too large number of binary variables and renders the approach intractable even for moderate-sized data sets. We have thus proposed another way of making a decision: let p_{ij} denote the partition with $n-1$ clusters, in which the only objects that are clustered together are objects i and j (partition p_{ij} is an atom in the lattice $(\Omega, \preceq)$). Then, the interval $[p_{ij}, p_E]$ represents the set of all partitions in which objects i and j are in the same cluster. Our belief in the fact that i and j belongs to the same cluster can be characterized by the credibility of $[p_{ij}, p_E]$, which can be computed as follows:

$$Bel_{ij} = bel([p_{ij}, p_E]) = \sum_{[\underline{p}_k, \overline{p}_k] \subseteq [p_{ij}, p_E]} m_k = \sum_{\underline{p}_k \succeq p_{ij}} m_k = \sum_{k=1}^{s} m_k \underline{R}_k(i,j). \qquad (27)$$

Matrix $Bel = (Bel_{ij})$ can be considered as a new similarity matrix and can in turn be clustered using, e.g., a hierarchical clustering algorithm. If a partition is needed, the classification tree (dendogram) can be cut at a specified level so as to insure a user-defined number of clusters.

Example 6 The data set used to illustrate the method is the half-ring data set inspired from [13]. It consists of two clusters of 100 points each in a two-dimensional space. To build the ensemble, we used the fuzzy c-means algorithm with a varying number of clusters (from 6 to 11). The six hard partitions computed from the soft partitions are represented in Figure 6.

Each hard partition p_ℓ ($\ell = 1, \ldots, 6$) was characterized by a confidence degree α_ℓ, which was computed using a validity index measuring the quality of the partition. Considering that the true partition is coarser than each individual one, and taking into account the uncertainty

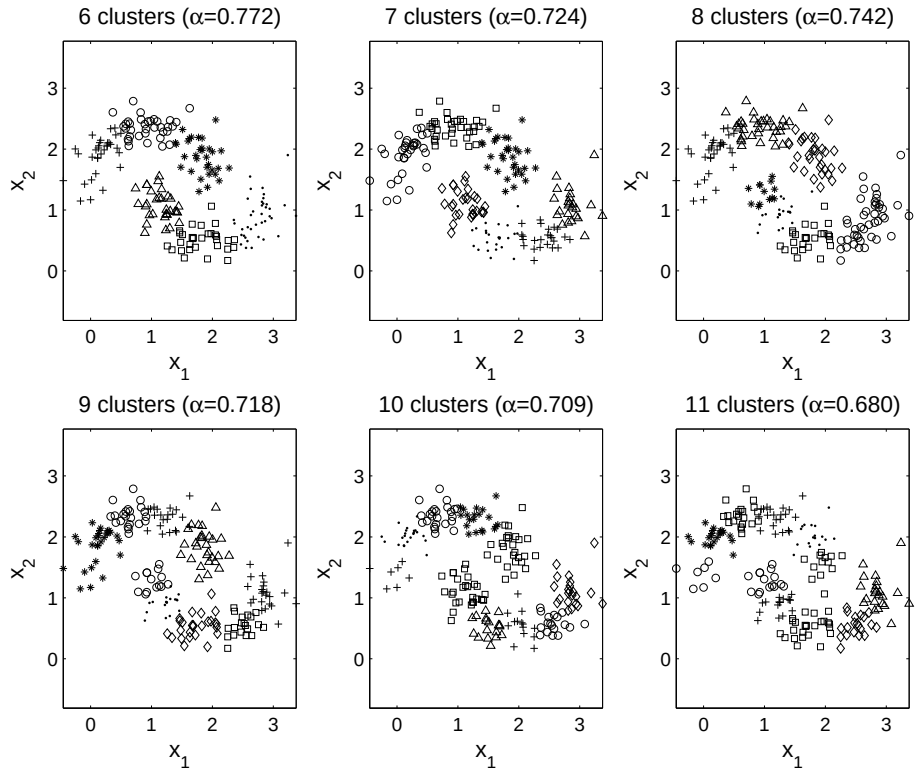

Fig. 6 Half-rings data set. Individual partitions.

of the clustering process, the following mass functions were defined:

$$\begin{cases} m^{\ell}([p_{\ell}, p_E]) = \alpha_{\ell} \\ m^{\ell}([p_0, p_E]) = 1 - \alpha_{\ell}. \end{cases} \tag{28}$$

The six mass functions (with two focal elements each) were then combined using Dempster's rule. A hierarchy was computed from matrix *Bel* using Ward's linkage. The dendrogram, represented in the left part of Figure 7, reveals a clear separation in two clusters. Cutting the tree to obtain two clusters yields the partition represented in the right part of Figure 7. We can see that the natural structure of the data is perfectly recovered.

7 Beliefs on Preorders

In Section 6.1, we presented the lattice of partitions of a finite set, which is isomorphic to the lattice of equivalence relations. We will now introduce a more general lattice of relations, the lattice of preorders, and we will illustrate its interest for aggregating preference relations.

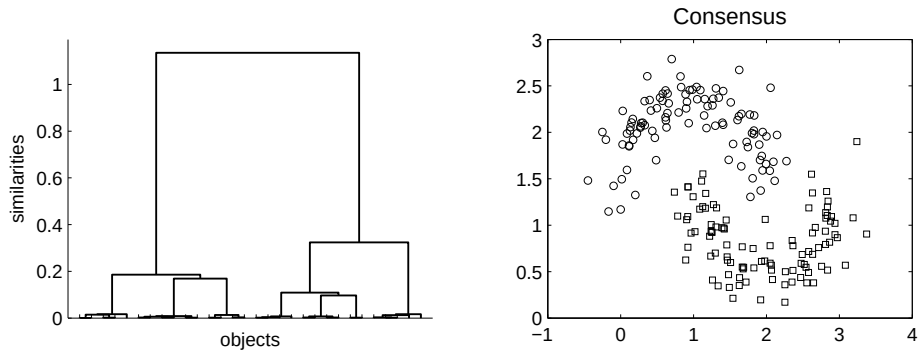

Fig. 7 Half-rings data set. Ward's linkage computed from *Bel* and derived consensus.

7.1 Lattice of preorders

Let $E = \{a_1, \ldots, a_n\}$ be a finite set of n objects. A *preorder* R on E is a binary reflexive and transitive relation. Preorders are very general relations encompassing partial orders (antisymmetric preorders) and equivalence relations (symmetric preorders) as special cases.

The set Ω of all preorders defined on E can be equipped with the same ordering relation as the one introduced for equivalence relations: a preorder R is said to be *finer* than a preorder R' (we write $R \preceq R'$) if and only if:

$$R(a_i, a_j) \leq R'(a_i, a_j) \quad \forall (a_i, a_j) \in E^2.$$

$(\Omega, \preceq)$ is a lattice. Note that the lattice of equivalence relations is a sublattice of the lattice of preorders. As for equivalence relations, the minimal element $\perp$ in this lattice is represented by a diagonal matrix and the maximal element $\top$ by a matrix with all entries equal to one. As in Section 6.1, the meet and the join of two elements R_1 and R_2 are defined, respectively, as the minimum of R_1 and R_2, and as the transitive closure of the maximum of R_1 and R_2. For any two relations $\underline{R}$ and $\overline{R}$ such that $\underline{R} \preceq \overline{R}$, we can define the interval $[\underline{R}, \overline{R}]$, i.e., the set of all preorders coarser than $\underline{R}$ and finer than $\overline{R}$:

$$[\underline{R}, \overline{R}] = \{R \in \Omega | \underline{R} \preceq R \preceq \overline{R}\}.$$

The set of intervals of the lattice $(\Omega, \preceq)$ defined by:

$$\mathscr{I}_{\Omega, \preceq} = \{[\underline{R}, \overline{R}] | \underline{R}, \overline{R} \in \Omega, \underline{R} \preceq \overline{R}\} \cup \emptyset_\Omega,$$

is also a lattice, endowed with the inclusion relation. This framework will be used in the next section for representing and aggregating partial information regarding the ranking of a set of objects.

7.2 Application to Preference Aggregation

It is often desirable, for decision making purposes, to define an order between different alternatives or objects, on the basis of different opinions or preferences provided by decision-makers. This problem has attracted considerable interest (see, e.g., [23, 25, 1, 15, 2]). Preferences are often expressed through pairwise comparisons. As explained in [24], comparing

two objects a_j and a_i can be seen as determining which of the four following possible situations holds:

1. Object a_i is "before" object a_j, where "before" implies some kind of order between a_i and a_j, induced either by direct preference (a_i is preferred to a_j) or by measurements on an associated scale;
2. The converse situation: object a_j is "before" object a_i;
3. Object a_i is "near" object a_j, where "near" can be considered either as indifference (object a_i or object a_j will do equally well for some purpose), or as a similarity;
4. Objects a_i and a_j cannot be compared because, for example, of lack of information, high uncertainty, or conflicting information.

A preference relation $R = (r_{ij})$ defined on the set of alternatives E is a preorder on E, which encodes the four above situations as follows:

1. Object a_i is strictly preferred to a_j ($a_i > a_j$) $\Leftrightarrow r_{ij} = 1$ and $r_{ji} = 0$;
2. Object a_j is strictly preferred to a_i ($a_i < a_j$) $\Leftrightarrow r_{ij} = 0$ and $r_{ji} = 1$;
3. Objects a_i et a_j are indifferent ($a_i \sim a_j$) $\Leftrightarrow r_{ij} = 1$ and $r_{ji} = 1$;
4. Objects a_i et a_j are incomparable ($a_i <> a_j$) $\Leftrightarrow r_{ij} = 0$ and $r_{ji} = 0$;

The problem addressed here is to derive such a preference relation from the opinions expressed by K agents about pairs of objects. More precisely, we suppose that, for each pair of alternatives $\{a_i, a_j\}$, each agent k has to choose between the four situations described above, and that he is able to quantify his degree of belief in the selected proposition by a value $\alpha_{ijk} \in [0,1]$. The choice of agent k for any pair $\{a_i, a_j\}$ with $i < j$ will be indicated using four binary variables $\delta_{ijk}^{(\ell)}$, $\ell = 1,4$, using the following convention:

$$\delta_{ijk}^{(1)} = \begin{cases} 1 \text{ if } a_i > a_j \\ 0 \text{ otherwise,} \end{cases} \tag{29}$$

$$\delta_{ijk}^{(2)} = \begin{cases} 1 \text{ if } a_i < a_j \\ 0 \text{ otherwise,} \end{cases} \tag{30}$$

$$\delta_{ijk}^{(3)} = \begin{cases} 1 \text{ if } a_i \sim a_j \\ 0 \text{ otherwise,} \end{cases} \tag{31}$$

$$\delta_{ijk}^{(4)} = \begin{cases} 1 \text{ if } a_i <> a_j \\ 0 \text{ otherwise.} \end{cases} \tag{32}$$

It is clear that exactly one of these four binary variables is equal to one, i.e., the following equation holds:

$$\sum_{\ell=1}^{4} \delta_{ijk}^{(\ell)} = 1.$$

Let us assume that an agent k has expressed his preference about objects a_i and a_j by setting the values of the four binary variables $\delta_{ijk}^{(\ell)}$, $\ell = 1, \ldots, 4$. This piece of information can be seen as defining a constraint on the underlying preorder R over the set E of objects. To express the set of preorders R compatible with this information, let us introduce two preorders ρ_{ij} and $\overline{\rho_{ij}}$ on E defined as follows:

$$\begin{cases} \rho_{ij}(a_i, a_j) = 1, \\ \rho_{ij}(a_k, a_l) = 0 \quad \forall (k,l) \neq (i,j). \end{cases}$$

and

$$\begin{cases} \overline{\rho_{ij}}(a_i,a_j)=0, \\ \overline{\rho_{ij}}(a_k,a_l)=1 \quad \forall (k,l)\neq(i,j). \end{cases}$$

It is clear that the following equivalences hold:

$$\begin{aligned} \delta_{ijk}^{(1)}=1 &\Leftrightarrow R\in \mathrm{I}_{ij}^{(1)}=[\rho_{ij},\overline{\rho_{ji}}], \\ \delta_{ijk}^{(2)}=1 &\Leftrightarrow R\in \mathrm{I}_{ij}^{(2)}=[\rho_{ji},\overline{\rho_{ij}}], \\ \delta_{ijk}^{(3)}=1 &\Leftrightarrow R\in \mathrm{I}_{ij}^{(3)}=[\rho_{ij}\vee\rho_{ji},\top], \\ \delta_{ijk}^{(4)}=1 &\Leftrightarrow R\in \mathrm{I}_{ij}^{(4)}=[\bot,\overline{\rho_{ij}}\wedge\overline{\rho_{ji}}]. \end{aligned}$$

Example 7 Let E be a set composed of three elements. Suppose that an agent has chosen the proposition $a_1 > a_2$. Then relation R has to belong to the following interval:

$$\left[\begin{pmatrix}1&1&0\\0&1&0\\0&0&1\end{pmatrix},\begin{pmatrix}1&1&1\\0&1&1\\1&1&1\end{pmatrix}\right]=[\rho_{12},\overline{\rho_{21}}],$$

i.e., all terms of matrix R are free except $r_{12}=1$ and $r_{21}=0$.

If the agent has chosen the indifference between a_1 and a_2, then the relation R is supposed to belong to:

$$\left[\begin{pmatrix}1&1&0\\1&1&0\\0&0&1\end{pmatrix},\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix}\right]=[\rho_{12}\vee\rho_{21};\top].$$

In this case, all terms of matrix R are free except $r_{12}=r_{21}=1$. □

The opinion of an agent k may be thus expressed in the lattice $(\mathscr{I}_{\Omega,\preceq},\subseteq)$ of preorder intervals by the following mass assignment (for all $i<j$):

$$\begin{cases} m_{ijk}(\mathrm{I}_{ij}^{(\ell)})=\delta_{ijk}^{(\ell)}\alpha_{ijk}, \quad \ell=1,\ldots,4, \\ m_{ijk}([\bot,\top])=1-\alpha_{ijk}. \end{cases} \tag{33}$$

The opinion given by K evaluators about the $n(n-1)/2$ pairs of objects may finally be combined into a single mass function m using Dempster's rule:

$$m=\bigoplus_{k=1}^{K}\bigoplus_{i<j} m_{ijk}. \tag{34}$$

Remark 1 The applicability of Dempster's rule is conditioned by the assumption of independence of the sources to be combined. The use of this rule can thus be questioned when combining several pairwise evaluations of a single agent. However, the dependency of these evaluations is not so straightforward, as explained in [33]. Even using unblinded comparisons, when comparing a_1 vs. a_2 and a_2 vs. a_3, the fact that a_2 is common to both comparisons does not necessarily imply a dependence of the evaluations, unless the expert is employing a notion of transitivity to force his choices to be coherent. The interested reader can refer to [33] for a detailed discussion about this issue. If the independence cannot be assumed, other combinations rules, such as the cautious conjunctive introduced [10], can be used. We may note here that this rule is based on the canonical decomposition of belief functions, which can be computed in any lattice [16]: consequently, it is well defined in the setting considered in this paper (see also [11]).

The corresponding commonality function q can be computed using (9) in the lattice $(\mathscr{I}_{\Omega,\prec},\subseteq)$ as:

$$q \propto \prod_{k=1}^{K} \prod_{i<j} q_{ijk}, \tag{35}$$

where q_{ijk} is the commonality function associated to m_{ijk}.

The commonality of a singleton $\{R\}$ is equal to:

$$q(\{R\}) \propto \prod_{k} \prod_{i<j} q_{ijk}(\{R\}), \tag{36}$$

with:

$$q_{ijk}(\{R\}) = \begin{cases} 1 & \text{if } \exists \ell \text{ such that } \delta_{ijk}^{(\ell)} = 1 \text{ and } R \in \mathrm{I}_{ij}^{(\ell)}, \\ 1-\alpha_{ijk} & \text{otherwise.} \end{cases} \tag{37}$$

Equation (37) can be simplified by encoding relation R using new variables $X_{ij}^{(\ell)}$ for $i<j$ and $\ell = 1,\ldots,4$, defined as follows:

$$\begin{cases} X_{ij}^{(1)} = r_{ij}(1-r_{ji}), \\ X_{ij}^{(2)} = r_{ji}(1-r_{ij}), \\ X_{ij}^{(3)} = r_{ij}r_{ji}, \\ X_{ij}^{(4)} = (1-r_{ij})(1-r_{ji}). \end{cases} \tag{38}$$

We thus have:

$$\begin{aligned} R \in \mathrm{I}_{ij}^{(1)} &\Leftrightarrow r_{ij} = 1 \text{ and } r_{ji} = 0 \Leftrightarrow X_{ij}^{(1)} = 1, \\ R \in \mathrm{I}_{ij}^{(2)} &\Leftrightarrow r_{ij} = 0 \text{ and } r_{ji} = 1 \Leftrightarrow X_{ij}^{(2)} = 1, \\ R \in \mathrm{I}_{ij}^{(3)} &\Leftrightarrow r_{ij} = 1 \text{ and } r_{ji} = 1 \Leftrightarrow X_{ij}^{(3)} = 1, \\ R \in \mathrm{I}_{ij}^{(4)} &\Leftrightarrow r_{ij} = 0 \text{ and } r_{ji} = 0 \Leftrightarrow X_{ij}^{(4)} = 1. \end{aligned}$$

With this notation, (37) can be rewritten as:

$$q_{ijk}(\{R\}) = \begin{cases} 1 & \text{if } \sum_{\ell=1}^{4} \delta_{ijk}^{(\ell)} X_{ij}^{(\ell)} = 1, \\ 1-\alpha_{ijk} & \text{otherwise.} \end{cases} \tag{39}$$

Maximizing $q(\{R\})$ is equivalent to maximizing its logarithm, which is equal to:

$$\ln q(\{R\}) = \sum_{k} \sum_{i<j} \left(1 - \sum_{\ell=1}^{4} \delta_{ijk}^{(\ell)} X_{ij}^{(\ell)}\right) \ln(1-\alpha_{ijk}) + \text{constant}. \tag{40}$$

By eliminating the constant term and permuting the sums, we can see that the optimal relation R maximizing (40) can finally be obtained by solving the following binary integer programming problem:

$$\min_{X_{ij}^{(\ell)} \in \{0,1\}} \sum_{i<j} \sum_{\ell=1}^{4} X_{ij}^{(\ell)} \sum_{k} \delta_{ijk}^{(\ell)} \ln(1-\alpha_{ijk}) \tag{41}$$

subject to the constraints:

$$\sum_{\ell=1}^{4} X_{ij}^{(\ell)} = 1, \quad \forall i < j \tag{42}$$

$$r_{ij} + r_{jk} - 1 \leq r_{ik} \quad \forall i \neq j \neq k. \tag{43}$$

Constraints (42) ensure that only one $X_{ij}^{(\ell)}$ is equal to 1. Constraints (43) ensure the transitivity of the resulting matrix R. Note that constraints (43) can be easily expressed with the unknowns of the problem, since the following relation holds:

$$r_{ij} = \begin{cases} X_{ij}^{(1)} + X_{ij}^{(3)} & \text{if } i < j, \\ X_{ij}^{(2)} + X_{ij}^{(3)} & \text{if } i > j. \end{cases} \tag{44}$$

Example 8 We illustrate our approach using a synthetic example. We generated at random a preorder $R^* = (r_{ij}^*)$ on a set E of $n = 6$ objects. Then, we assumed that a varying number of respondents (from 1 to 35) were asked for their opinion about the 15 pairs of objects. Their evaluation was simulated as follows. For each pair $\{a_i, a_j\}$ and each respondent k, a probability p_{ijk} of error was drawn from a beta distribution. Then, the true evaluation was kept with probability $1 - p_{ijk}$, and changed with probability p_{ijk}, a wrong evaluation being uniformly chosen among the three remaining possibilities. The parameters of the beta distribution were varied so as to induce different noise levels in the evaluations (the first parameter of the distribution was set to 20 and the second parameter was chosen in the set $\{0.5, 2, 5, 10, 15, 18, 20\}$ so that the expectation for the error probability varied between 2% and 50%). Lastly, we assumed that each respondent was able to quantify the error level in his evaluations, so that we fixed $\alpha_{ijk} = 1 - p_{ijk}$, $\forall i < j$ and $\forall k$. The relation R obtained by solving problem (41-43) was compared to the true preorder using the error rate defined as:

$$e = \frac{2}{n(n-1)} \sum_{i<j} \mathbb{1}_{r_{ij} \neq r_{ij}^*}$$

The simulation process was repeated 100 times and the results were averaged. They are presented in Figure 8. We can see from the figure that our approach acts as a good denoising process, since, for every noise levels, the error tends to zero as the number of respondents grows. Note that a reduction of error rate is also observed even if only one respondent is considered.

When pooling the opinions of the experts using the linear programming approach (41), it is clear that the internal coherence of each expert is not taken into account. If one wants to take into account the self consistency of each expert, it is possible to use a two-step procedure. The first step consists in assessing the internal coherence of each expert by applying the decision making approach to each preorder and discounting the opinion of the experts according to this coherence. In fact, when the procedure is applied individually to each expert, the maximal value of $q(\{R\})$ reflects the consistency of the expert: if this value is equal to 1, then the evaluations of the expert are fully consistent. A value less than 1 reflects some conflict in the evaluations of the expert. The maximal commonality values for each expert can thus be used to discount expert opinions, before applying the overall procedure.

Example 9 We illustrate this point using an example inspired from [33]. In a study conducted at the Ontario Cancer Institute, subjects were asked to give their preferences about four scenarios describing ethical dilemmas in health care. The preferences for all six possible scenario pairs were obtained. The experts were also asked to rate the reliability of their evaluations. The preferences for one subject (expert 1) were:

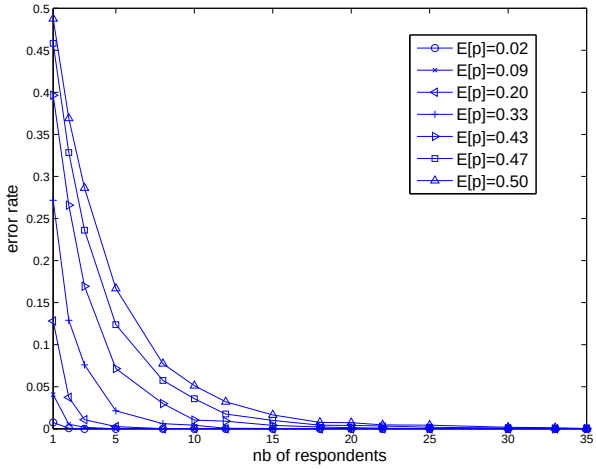

Fig. 8 Error rate as a function of the number of respondents.

(a) (b)

Fig. 9 Graph representation of the evaluations; (a): expert 1 ($\max q(\{R\}) = 0.56$); (b): expert 2 ($\max q(\{R\}) = 1$).

- Scenario B is preferred to scenario A with reliability 0.44,
- Scenario D is preferred to scenario B with reliability 0.74,
- Scenario D is preferred to scenario C with reliability 0.06,
- Scenario A is preferred to scenario C with reliability 0.97,
- Scenario A is preferred to scenario D with reliability 0.94,
- Scenario B is preferred to scenario C with reliability 0.93.

The preference of a subject can be represented by a directed graph in which the vertices are the scenarios and the edges represent the relation “is preferred to”. The corresponding graph of expert 1 is given in Figure 9a. The fact that the graph contains a cycle (ADB) shows that the evaluations of expert 1 are not fully consistent. We suppose now that the evaluations of a second subject (expert 2), represented in Figure 9b, are given. This time, there is no cycle in the graph, but the degrees of belief are weaker than for expert 1. Applied individually to each expert, the procedure gives a a commonality of 0.56 for the first expert, and a commonality of 1 for the second expert. If the evaluations are merged directly, one obtains the relation represented in Figure 10a, which is close to the relation given by expert 1 with a correction of transitivity. If the evaluations of the experts are beforehand discounted using one minus the individual commonalities as discounting factors, then the relation represented in Figure 10b with a commonality equal to 0.36 is found, which is the relation given by expert 2.

(a) (b)

Fig. 10 Graph representation of the aggregation; (a): without discounting ($\max q(\{R\}) = 0.11$); (b): with discounting ($\max q(\{R\}) = 0.29$).

8 Conclusion

The exponential complexity of operations in the theory of belief functions has long been seen as a shortcoming of this approach, and has prevented its application to very large frames of discernment. We have shown in this paper that the complexity of the Dempster-Shafer calculus can be drastically reduced, while retaining sufficient expressive power, if belief functions are defined over a suitable subset of the power set equipped with a lattice structure. When the frame of discernment forms itself a lattice for some partial ordering, the set of events may be defined as the set of intervals in that lattice. Using this method, it is possible to define and manipulate belief functions in very large frames such as the power set of a finite set, the set of partitions of n objects or the set of preorders over a set of alternatives. This approach opens the way to the application of Dempster-Shafer theory to a wide range of computationally demanding tasks in Decision Analysis and Machine Learning.

References

1. I. Charon and O. Hudry. Lamarckian genetic algorithms applied to the aggregation of preferences. *Annals of Operations Research*, 80:281–297, 1998.
2. I. Charon and O. Hudry. An updated survey on the linear ordering problem for weighted or unweighted tournaments. *Annals of Operations Research*, 175:107–158, 2010.
3. B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–330, 2006.
4. A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.
5. A. P. Dempster. Upper and lower probabilities generated by a random closed interval. *Annals of Mathematical Statistics*, 39(3):957–966, 1968.
6. T. Denœux. A *k*-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25(05):804–813, 1995.
7. T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.
8. T. Denœux. Inner and outer approximation of belief structures using a hierarchical clustering approach. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(4):437–460, 2001.
9. T. Denœux. Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3):228–252, 2006.
10. T. Denoeux. Conjunctive and Disjunctive Combination of Belief Functions Induced by Non Distinct Bodies of Evidence. *Artificial Intelligence*, 172:234–264, 2008.
11. T. Denoeux, Z. Younes, and F. Abdallah. Representing uncertainty on set-valued variables using belief functions. *Artificial Intelligence*, 174(7-8):479–499, 2010.
12. D. Dubois and H. Prade. A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. *International Journal of General Systems*, 12(3):193–226, 1986.

13. A. Fred and A. Jain. Data clustering using evidence accumulation. In *Proceedings of the 16th International Conference on Pattern Recognition*, pages 276–28, Quebec, Canada, 2002.
14. A. Fred and A. Lourenço. Cluster ensemble methods: from single clusterings to combined solutions. *Studies in Computational Intelligence (SCI)*, 126:3–30, 2008.
15. J. Gonzales-Pachon and C. Romero. Inferring consensus weights from pairwise comparison matrices without suitable properties. *Annals of Operations Research*, 154:123–132, 2007.
16. M. Grabisch. Belief functions on lattices. *International Journal of Intelligent Systems*, 24:76–95, 2009.
17. M. Grabisch and C. Labreuche. Bi-capacities – I: definition, Möbius transform and interaction. *Fuzzy Sets and Systems*, 151:211–236, 2005.
18. K.Hornik and F. Leisch. Ensemble methods for cluster analysis. In A. Taudes, editor, *Adaptive Information Systems and Modelling in Economics and Management Science, Volume 5 of Interdisciplinary Studies in Economics and Management*, pages 261–268. Springer-Verlag, 2005.
19. L. Liu. A theory of Gaussian belief functions. *International Journal of Approximate Reasoning*, 14:95–126, 1996.
20. M.-H. Masson and T. Denœux. Belief functions and cluster ensembles. In C. Sossai and G. Chemello, editors, *Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2009)*, pages 323–334, Verona, Italy, 2009. Springer-Verlag.
21. M.-H. Masson and T. Denœux. Ensemble clustering in the belief functions framework. *International Journal of Approximate Reasoning*, 52(1):92-109, 2011.
22. B. Monjardet. The presence of lattice theory in discrete problems of mathematical social sciences. Why. *Mathematical Social Sciences*, 46(2):103–144, 2003.
23. J. W. Owsinski. Preferences, agreement, consensus - measuring, aggregation and control. Annals of Operations Research, 51(5):217–240, 1994.
24. M. Oztürk, A. Tsoukiàs, and P. Vincke. Preference modelling. In J. Figueira, S. Greco, and M. Ehrgott, editors, *Multiple Criteria Decision Analysis: State of the Art Surveys*, pages 27–72. Springer Verlag, Boston, Dordrecht, London, 2005.
25. P. Perny. Multicriteria filtering methods based onconcordance and non-discordance principles. Annals of Operations Research, 80(0):137-165, 1998.
26. G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.
27. G. Shafer, P. P. Shenoy, and K. Mellouli. Propagating belief functions in qualitative Markov trees. *International Journal of Approximate Reasoning*, 1:349–400, 1987.
28. P. P. Shenoy. Binary joint trees for computing marginals in the Shenoy-Shafer architecture. *International Journal of Approximate Reasoning*, 17:239–263, 1997.
29. P. Smets. The combination of evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458, 1990.
30. P. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.
31. P. Smets. Belief functions on real numbers. *International Journal of Approximate Reasoning*, 40(3):181–223, 2005.
32. P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.
33. D. Tritchler and G. Lockwood. Modelling the reliability of paired comparisons. *Journal of Mathematical Psychology*, 35:277–293, 1991.
34. K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multi-label classification of music into emotions. In *Proc. 9th International Conference on Music Information Retrieval (ISMIR 2008)*, Philadephia, PA, USA, 2008.
35. N. Wilson. Algorithms for Dempster-Shafer theory. In D. M. Gabbay and P. Smets, editors, *Handbook of defeasible reasoning and uncertainty management. Volume 5: Algorithms for Uncertainty and Defeasible Reasoning*, pages 421–475. Kluwer Academic Publishers, Boston, 2000.
36. R. R. Yager. Set-based representations of conjunctive and disjunctive knowledge. *Information Sciences*, 41:1–22, 1987.
37. Z. Younes, F. Abdallah, and T. Denœux. An evidence-theoretic k-nearest neighbor rule for multi-label classification. In *Proceedings of the 3rd International Conference on Scalable Uncertainty Management (SUM 2009)*, number 5785 in LNAI, pages 297–308, Washington, DC, USA, 2009. Springer-Verlag.
38. Z. Younes, F. Abdallah, and T. Denoeux. Evidential multi-label classification approach to learning from data with imprecise labels. In *Proceedings of IPMU 2010*, Dortmund, Germany, July 2010.
39. M.-L. Zhang and Z.-H. Zhou. ML-KNN: a lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.