Representing uncertainty on set-valued variables using belief functions

Thierry Denœux[1], Zoulficar Younes, Fahed Abdallah

HEUDIASYC, UTC, CNRS
Centre de Recherche de Royallieu
BP 20529, F-60205 Compiègne, France

February 2, 2010

[1]Corresponding author. Email address: tdenoeux@hds.utc.fr. Fax: +33 (0)3 44 23 44 77.

Abstract

A formalism is proposed for representing uncertain information on set-valued variables using the formalism of belief functions. A set-valued variable X on a domain Ω is a variable taking zero, one or several values in Ω. While defining mass functions on the frame 2^{2^Ω} is usually not feasible because of the double-exponential complexity involved, we propose an approach based on a definition of a restricted family of subsets of 2^Ω that is closed under intersection and has a lattice structure. Using recent results about belief functions on lattices, we show that most notions from Dempster-Shafer theory can be transposed to that particular lattice, making it possible to express rich knowledge about X with only limited additional complexity as compared to the single-valued case. An application to multi-label classification (in which each learning instance can belong to several classes simultaneously) is demonstrated.

Keywords: Dempster-Shafer theory, Evidence theory, conjunctive knowledge, lattice, uncertain reasoning, multi-label classification.

1 Introduction

An important concept in knowledge representation is that of *variable*. Usually, we associate to each variable X a *domain* (or *frame of discernment*) Ω, and we assume that X takes one and only one value in Ω. For instance, in conventional classification problems, X denotes the class of an object, and each object is assumed to belong to one and only one class among a set Ω of classes.

There are cases, however, where it is convenient to consider a variable X taking zero, one or several values in a domain Ω. In such cases, X may be called a *set-valued*, or *conjunctive* variable [8, 33]. For instance, in diagnosis problems, Ω may denote the set of faults that can possibly occur in a system, and X the faults actually occurring at a given time. In text classification, Ω may denote a set of topics, and X the list of topics dealt with in a given text, etc.

A straightforward approach to the above problem is, of course, to consider a set-valued variable X on Ω as a single-valued variable on the power set $\Theta = 2^\Omega$. However, this approach often implies working in a space of very high cardinality. If, as done in this paper, we assume Ω to be finite with size K, then the size of Θ is 2^K. If we want to express imprecise information about X, we will have to manipulate subsets of Θ. As there are 2^{2^K} of these subsets, this approach rapidly becomes intractable as K increases.

In this paper, we consider the problem of representing partial knowledge about a set-valued variable X with domain Ω using the Dempster-Shafer theory of belief functions [26, 30]. Our approach will be based on a simple representation of a class $\mathcal{C}(\Omega)$ of subsets of $\Theta = 2^\Omega$ which, endowed with set inclusion, has a lattice structure. Using recent results about belief functions on lattices [14], we will be able to generalize most concepts of Dempster-Shafer theory (including the canonical decompositions and the cautious rule [5]) in this setting. This formalism will be shown to allow the expression of a wide range of knowledge about set-valued variables, with only a moderate increase of complexity (from 2^K to 3^K) as compared to the usual single-valued case.

The rest of this paper is organized as follows. Background notions on belief functions in the classical setting and in general lattices will first be recalled in Sections 2 and 3, respectively. Our approach will then be introduced in Section 4, and some relationships with previous work will be outlined in Section 5. An application to multi-label classification

will be presented in Section 6, and Section 7 will conclude the paper.

2 Belief Functions

The basic concepts of the Dempster-Shafer theory of belief functions, as introduced in [26], will first be summarized in Subsection 2.1. The canonical decomposition and the cautious rule will then be recalled in Section 2.2.

2.1 Basic definitions

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

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

The subsets A of Ω such that $m(A) > 0$ are called the *focal elements* of m. The set of focal elements of m will be denoted $\mathcal{F}(m)$. m is said to be *normal* if $\emptyset$ is not a focal set, and *dogmatic* if Ω is not a focal set.

A mass function m is often used to model an agent's beliefs about a variable X taking a single but ill-known value ω_0 in Ω [30]. 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 $m(\Omega) = 1$. Probabilistic uncertainty corresponds to the case where all focal elements are singletons, in which case m is equivalent to a probability distribution on Ω.

To each mass function m can be associated an *implicability function* b and a *belief function bel* defined as follows:

$$b(A) = \sum_{B \subseteq A} m(B) \tag{1}$$

$$bel(A) = \sum_{B \subseteq A, B \not\subseteq \overline{A}} m(B) = b(A) - m(\emptyset). \tag{2}$$

These two functions are equal when m is normal. However, they need to be distinguished when considering non normal mass functions. Function *bel* has easier interpretation, as $bel(A)$ corresponds to a degree of belief in the proposition "The true value ω_0 of X belongs

to A". However, function b has simpler mathematical properties. For instance, m can be recovered from b as

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

where $|\cdot|$ denotes cardinality. Function m is said to be the *Möbius transform* of b. For every function f from 2^Ω to $[0,1]$ such that $f(\Omega) = 1$, 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, b (and *bel*) are totally mononotone.

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

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \tag{4}$$

$$= 1 - b(\overline{A}) \tag{5}$$

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

$$q(A) = \sum_{B \supseteq A} m(B). \tag{6}$$

m can be recovered from q using the following relation:

$$m(A) = \sum_{B \supseteq A} (-1)^{|B \setminus A|} q(B). \tag{7}$$

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

A special case of interest is that where the focal elements of m are nested: m is then said to be *consonant*. In this case, we have

$$pl(A \cup B) = \max(pl(A), pl(B)), \quad \forall A, B \subseteq \Omega.$$

The plausibility function is thus a possibility measure, with corresponding possibility distribution defined by $\pi(x) = pl(\{x\})$ for all $x \in \Omega$. Conversely, to each possibility distribution corresponds a unique consonant mass function [26].

Let us now assume that we receive two 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 the *conjunctive sum* (or unnormalized Dempster's rule of combination) defined as follows:

$$(m_1 \textcircled{\cap} m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C). \tag{8}$$

This rule is commutative, associative, and admits the vacuous mass function as neutral element. It is conjunctive as the product of $m_1(B)$ and $m_2(C)$ is transferred to the intersection of B and C. The quantity $(m_1 \textcircled{\cap} m_2)(\emptyset)$ is referred to as the *degree of conflict* between m_1 and m_2.

Let $q_{1\textcircled{\cap}2}$ denote the commonality function corresponding to $m_1 \textcircled{\cap} m_2$. It can be computed from q_1 and q_2, the commonality functions associated to m_1 and m_2, as follows:

$$q_{1\textcircled{\cap}2}(A) = q_1(A) \cdot q_2(A), \quad \forall A \subseteq \Omega. \tag{9}$$

The normalized Dempster's rule $\oplus$ [26] is defined as the conjunctive sum followed by a normalization step:

$$(m_1 \oplus m_2)(A) = \begin{cases} 0 & \text{if } A = \emptyset \\ \frac{(m_1 \textcircled{\cap} m_2)(A)}{1 - (m_1 \textcircled{\cap} m_2)(\emptyset)} & \text{otherwise.} \end{cases} \tag{10}$$

It is clear that $m_1 \oplus m_2$ is defined as long as $(m_1 \textcircled{\cap} m_2)(\emptyset) < 1$.

Alternatives to the conjunctive sum can be constructed by replacing $\cap$ by any binary set operation in (8). For instance, the choice of the union operator results in the *disjunctive sum* [28]:

$$(m_1 \textcircled{\cup} m_2)(A) = \sum_{B \cup C = A} m_1(B) m_2(C). \tag{11}$$

It can be shown that

$$b_{1\textcircled{\cup}2}(A) = b_1(A) \cdot b_2(A), \quad \forall A \subseteq \Omega, \tag{12}$$

which is the counterpart of (9). Dubois and Prade [10] have also proposed a "hybrid" rule intermediate between the conjunctive and disjunctive sums, in which the product $m_1(B)m_2(C)$ is assigned to $B \cap C$ whenever $B \cap C \neq \emptyset$, and to $B \cup C$ otherwise. This rule is not associative, but it usually provides a good summary of partially conflicting items of evidence.

In [30], Smets proposed a two-level model in which items of evidence are quantified by mass functions and combined at the *credal* level, while decisions are made at the *pignistic* level (from the Latin *pignus* meaning a bet). Once a decision has to be made, a mass function m is thus transformed into a *pignistic probability distribution* p. The pignistic transformation consists in normalizing m (assuming that $m(\emptyset) < 1$), and then distributing each normalized mass $m(A)/(1 - m(\emptyset))$ equally between the atoms $\omega_k \in A$:

$$p(\omega_k) = \sum_{\{A \subseteq \Omega, \omega_k \in A\}} \frac{m(A)}{(1 - m(\emptyset))|A|}, \quad \forall \omega_k \in \Omega. \tag{13}$$

Other authors have suggested the so-called plausibility transformation for transforming a mass function into a probability distribution, by normalizing the plausibilities of singletons [3]. In a decision making context, this approach results in selecting the most plausible single hypothesis.

2.2 Canonical Decompositions and Idempotent Rules

According to Shafer [26], a mass function is said to be *simple* if it has the following form

$$\begin{aligned} m(A) &= 1 - w_0 \\ m(\Omega) &= w_0, \end{aligned}$$

for some $A \subset \Omega$ and some $w_0 \in [0, 1]$. Let us denote such a mass function as A^{w_0}. The vacuous mass function may thus be noted A^1 for any $A \subset \Omega$. It is clear that

$$A^{w_0} \textcircled{\cap} A^{w_0'} = A^{w_0 w_0'}.$$

A mass function may be called *separable* if it can be obtained as the result of the conjunctive sum of simple mass functions. It can then be written:

$$m = \textcircled{\cap}_{A \subset \Omega} A^{w(A)}, \tag{14}$$

with $w(A) \in [0, 1]$ for all $A \subset \Omega$.

Smets [29] showed that any non dogmatic mass function m can be uniquely expressed using (14), with weights $w(A)$ now in $(0, +\infty)$. This is referred to as the *conjunctive canonical decomposition* of a mass function. Note that, when $w(A) > 1$, $A^{w(A)}$ is no longer a mass function, but the conjunctive sum can be extended to such "generalized mass functions" in an obvious way.

Function w is called the *conjunctive weight function* associated to m [5]. It is a new equivalent representation of a non dogmatic mass function, which may be computed directly from q as follows:

$$w(A) = \prod_{B \supseteq A} q(B)^{(-1)^{|B \setminus A|+1}}, \quad \forall A \subset \Omega, \tag{15}$$

or, taking logarithms,

$$\ln w(A) = -\sum_{B \supseteq A} (-1)^{|B \setminus A|} \ln q(B), \quad \forall A \subset \Omega. \tag{16}$$

In [29] and [5], $w(A)$ was defined for all strict subsets A of Ω. However, function w can be extended to 2^Ω by using (15) for $A = \Omega$. We then have:

$$w(\Omega) = \frac{1}{q(\Omega)} = \frac{1}{m(\Omega)} = \left(\prod_{A \subset \Omega} w(A) \right)^{-1}$$

and

$$\prod_{A \subseteq \Omega} w(A) = 1. \tag{17}$$

With this convention, (16) can be extended to all $A \subseteq \Omega$. We notice that (16) then has exactly the same form as (7), i.e., the formula for computing $\ln w$ from $-\ln q$ is the same as the one for computing m from q. Conversely, $\ln q$ can thus be computed from $-\ln w$ using a formula similar to (6):

$$\ln q(A) = -\sum_{B \supseteq A} \ln w(B), \quad \forall A \subseteq \Omega.$$

We note that function w has a simple property with respect to the conjunctive sum. Let w_1 and w_2 be two weight functions, and let $w_{1 \textcircled{\cap} 2}$ denote the result of their $\textcircled{\cap}$-combination. Then the following relation holds:

$$w_{1 \textcircled{\cap} 2}(A) = w_1(A) w_2(A), \quad \forall A \subseteq \Omega. \tag{18}$$

In [5], Denœux introduced the *cautious rule*, noted $\textcircled{\wedge}$, which is obtained by replacing the product by the minimum in (18), for all $A \subset \Omega$:

$$w_{1 \textcircled{\wedge} 2}(A) = \min\left(w_1(A), w_2(A)\right). \tag{19}$$

The value of $w_{1 \textcircled{\wedge} 2}(\Omega)$ can then be determined to satisfy the normalization condition (17). This rule is obviously commutative, associative and idempotent. As shown in [5],

it is suitable for combining conjunctively non independent items of evidence. As the conjunctive sum, the cautious rule has a normalized version defined by

$$(m_1 \textcircled{\wedge}^* m_2)(A) = \begin{cases} 0 & \text{if } A = \emptyset \\ \frac{(m_1 \textcircled{\wedge} m_2)(A)}{1-(m_1 \textcircled{\wedge} m_2)(\emptyset)} & \text{otherwise.} \end{cases} \tag{20}$$

As shown in [5], the conjunctive canonical decomposition also has a disjunctive counterpart. Any mass function m such that $m(\emptyset) > 0$ can be decomposed disjunctively as follows:

$$m = \textcircled{\cup}_{A \supset \emptyset} A_{v(A)}, \tag{21}$$

where $A_{v(A)}$ is a generalized mass function assigning a mass $v(A) > 0$ (possibly greater than 1) to $\emptyset$, and $1 - v(A)$ to A, for all $A \subseteq \Omega$, $A \neq \emptyset$. This defines a new function v, called the *disjunctive weight function*, which can be computed from b as follows:

$$v(A) = \prod_{B \subseteq A} b(B)^{(-1)^{|A \setminus B|+1}}, \quad \forall A \subseteq \Omega, A \neq \emptyset, \tag{22}$$

or

$$\ln v(A) = -\sum_{B \subseteq A} (-1)^{|A \setminus B|} \ln b(B), \quad \forall A \subseteq \Omega, A \neq \emptyset. \tag{23}$$

As before, the above equations can be extended to $A = \emptyset$, which leads to

$$v(\emptyset) = \frac{1}{b(\emptyset)} = \frac{1}{m(\emptyset)} = \left(\prod_{A \neq \emptyset} v(A) \right)^{-1}$$

and

$$\prod_{A \subseteq \Omega} v(A) = 1. \tag{24}$$

The disjunctive rule (11) has a simple expression as a function of disjunctive weights:

$$v_{1 \textcircled{\cup} 2}(A) = v_1(A) v_2(A), \quad \forall A \subseteq \Omega. \tag{25}$$

By replacing the product by the minimum in the above equation, we can define a new rule, denoted $\textcircled{\vee}$ and called the *bold rule* in [5]:

$$v_{1 \textcircled{\vee} 2}(A) = \min\left(v_1(A), v_2(A)\right), \quad A \subseteq \Omega, A \neq \emptyset, \tag{26}$$

and $v_{1 \textcircled{\vee} 2}(\emptyset) = \left(\prod_{A \neq \emptyset} v_{1 \textcircled{\vee} 2}(A) \right)^{-1}$. This rule is obviously commutative, associative and idempotent; it is suitable for combining disjunctively non independent items of evidence.

3 Extension to General Lattices

As shown by Grabisch [14], the theory of belief function can be extended from the Boolean lattice $(2^\Omega, \subseteq)$ to any lattice, not necessarily Boolean. We will first recall some basic definitions about lattices in Section 3.1. Grabisch's results used in this work will then be summarized in Section 3.2.

3.1 Lattices

A review of lattice theory can be found in [21]. The following presentation follows [14].

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 (denote $\top$) and the least element (denoted $\bot$) always exist. We say that x *covers* y if $x > y$ and there is no z such that $x > z > y$. 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. 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 $\top = \Omega$.

A *closure system* $\mathcal{C}$ on a set Θ is a family of subsets of Θ satisfying the following properties:

1. $\Theta \in \mathcal{C}$.
2. $C_1, C_2 \in \mathcal{C} \Rightarrow C_1 \cap C_2 \in \mathcal{C}$.

As shown in [21], any closure system $(\mathcal{C}, \subseteq)$ is a lattice with the following meet and join

operations

$$C_1 \wedge C_2 = C_1 \cap C_2 \tag{27}$$

$$C_1 \vee C_2 = \bigcap\{C \in \mathcal{C} | C_1 \cup C_2 \subseteq C\}. \tag{28}$$

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:

$$f(x) = \sum_{y \leq x} m(y), \quad \forall x \in L. \tag{29}$$

Function m can be expressed as:

$$m(x) = \sum_{y \leq x} \mu(y, x) f(y), \tag{30}$$

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

$$\mu(x, y) = \begin{cases} 1 & \text{if } x = y, \\ -\sum_{x \leq t < y} \mu(x, t) & \text{if } x < y, \\ 0, & \text{otherwise.} \end{cases} \tag{31}$$

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

$$q(x) = \sum_{y \geq x} m(y), \tag{32}$$

and m can be recovered from q as:

$$m(x) = \sum_{y \geq x} \mu(x, y) q(y). \tag{33}$$

Let us now assume that $(L, \leq)$ is a lattice. Following Grabisch [14], a function $b : L \to [0, 1]$ will be called an implicability function on L if $b(\top) = 1$, and its Möbius transform is nonnegative. The corresponding belief function *bel* can then be defined as:

$$bel(x) = b(x) - m(\bot), \quad \forall x \in L.$$

Note that Grabisch [14] considered only normal belief functions, in which case $b = bel$. As shown in [14], any implicability function on $(L, \leq)$ is totally monotone, i.e., for any $k \geq 2$ and for any family $x_1, \ldots, x_k$ in L,

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

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

As shown in [14], most results of Dempster-Shafer theory can be transposed in the general setting of lattices. For instance, the conjunctive sum (8) becomes:

$$(m_1 \textcircled{\cap} m_2)(x) = \sum_{y \wedge z = x} m_1(y) m_2(z), \quad \forall x \in L, \tag{34}$$

and the following relation between commonality functions still holds:

$$q_{1\textcircled{\cap}2}(x) = q_1(x) \cdot q_2(x), \quad \forall x \in L. \tag{35}$$

The normalized Dempster's rule $\oplus$ can still be defined, as in the classical case, by dividing each number $(m_1 \textcircled{\cap} m_2)(x)$ with $x \neq \bot$ by $1 - (m_1 \textcircled{\cap} m_2)(\bot)$, provided that $(m_1 \textcircled{\cap} m_2)(\bot) < 1$.

Using a similar line of reasoning as that followed in [14], we can also extend the disjunctive rule (11) as:

$$(m_1 \textcircled{\cup} m_2)(x) = \sum_{y \vee z = x} m_1(y) m_2(z), \quad \forall x \in L, \tag{36}$$

and (12) becomes:

$$b_{1\textcircled{\cup}2}(x) = b_1(x) \cdot b_2(x), \quad \forall x \in L. \tag{37}$$

Grabisch [14] also extended the conjunctive canonical decomposition of belief functions in the general lattice setting. He showed that any mass function m on L such that $m(\top) > 0$ can be decomposed as

$$m = \textcircled{\cap}_{x < \top} x^{w(x)}, \tag{38}$$

where $x^{w(x)}$ is a simple mass function assigning $1 - w(x)$ to x and $w(x)$ to $\top$, with $w(x) \in (0, +\infty)$. Clearly, (38) generalizes (14). As in the classical case, function w :

$L \setminus \{\top\} \to (0, +\infty)$ can be computed from q using the following equation:

$$w(x) = \prod_{y \geq x} q(y)^{-\mu(x,y)}, \quad \forall x \in L, x \neq \top, \tag{39}$$

which generalizes (15). Obviously, we still have

$$w_{1 \textcircled{\cap} 2}(x) = w_1(x) w_2(x), \quad \forall x \in L, x \neq \top. \tag{40}$$

The existence of the w function also allows us to define the cautious rule in the general lattice setting as

$$w_{1 \textcircled{\wedge} 2}(x) = \min\left(w_1(x), w_2(x)\right), \quad \forall x \in L, x \neq \top. \tag{41}$$

The normalized cautious rule $\textcircled{\wedge}^*$ is defined as in the classical case, by dividing each $w_{1 \textcircled{\wedge} 2}(x)$ for $x \neq \bot$ by $1 - w_{1 \textcircled{\wedge} 2}(\bot)$, provided that $w_{1 \textcircled{\wedge} 2}(\bot) < 1$.

Although Grabisch did not consider the disjunctive canonical decomposition, it can also be extended in the general lattice setting. The proof parallels that given in [14] for the conjunctive case. We will only state the main result here. Let $x_{v(x)}$ be a mass function on L assigning $1 - v(x)$ to x and $v(x)$ to $\bot$, with $v(x) \in (0, +\infty)$. Any mass function m on L such that $m(\bot) > 0$ can be decomposed as

$$m = \textcircled{\cup}_{x > \bot} x_{v(x)}. \tag{42}$$

The function $v : L \setminus \{\bot\} \to (0, +\infty)$ can be computed from b using the following equation:

$$v(x) = \prod_{y \leq x} b(y)^{-\mu(y,x)}, \quad \forall x \in L, x \neq \bot, \tag{43}$$

which generalizes (22). We still have

$$v_{1 \textcircled{\cup} 2}(x) = v_1(x) v_2(x), \quad \forall x \in L, x \neq \bot, \tag{44}$$

and the existence of the v function allows us to define the bold rule as

$$v_{1 \textcircled{\vee} 2}(x) = \min\left(v_1(x), v_2(x)\right), \quad \forall x \in L, x \neq \bot. \tag{45}$$

The extension of other notions from classical Dempster-Shafer theory may require additional assumptions on $(L, \leq)$. For instance, the definition of the plausibility function pl as the dual of b using (5) can only be extended to autodual lattices [14]. The definition

of pl from (4) remains possible in the other cases, but the relationship between pl and b (or bel) is lost. Also, probability measures cannot be defined on arbitrary lattices. Consequently, the pignistic probability (13) can only be extended in restricted settings.

Remark 1 Although our approach relies essentially on Grabisch's work, we may note the existence of another line of research that aims at extending results of Probability Theory to some classes of residuated lattices, which are more general than Boolean algebra. In particular, there have been many developments about probability measures on MV-algebra (aso called *states*), see, e.g., [2, 16, 17, 22] as well as in Gödel algebras [1]. In the course of revising this paper, we also became aware of recent work on defining belief functions on MV-algebras [18][1].

4 Belief Functions on Set-valued Variables

In this section, the main concepts of Dempster-Shafer theory recalled in Section 2 will be extended to the case where we want to describe the uncertainty regarding a set-valued variable X on a finite domain Ω. The key to this extension will be the definition of a closure system $\mathcal{C}(\Omega)$ of $\Theta = 2^\Omega$, i.e., a set of subsets of Θ that is closed under intersection. Each element of $\mathcal{C}(\Omega)$ will be shown to have a simple description as a pair of disjoint subsets of Ω. Belief functions and associated notions will then be defined on the lattice $(\mathcal{C}(\Omega), \subseteq)$, resulting in a simple framework for uncertain reasoning about set-valued variables.

4.1 The Lattice $(\mathcal{C}(\Omega), \subseteq)$

In the rest of this paper, X denotes a set-valued variable on a finite domain Ω, i.e., a variable taking values in $\Theta = 2^\Omega$. Let $A_0 \subseteq \Omega$ denote the unknown true value of X. We want to describe partial knowledge about that value in the belief function framework.

As explained in the introduction, the formalism recalled in Section 2 could be applied without modification to this case, by defining a mass function m^Θ on Θ. However, such a brute force approach would require the storage of up to $2^{|\Theta|} = 2^{2^{|\Omega|}}$ numbers for each mass function. Basic operations such as the conjunctive or disjunctive sums would have

[1] We are indebted to one of the anonymous referees for bringing this work to our attention.

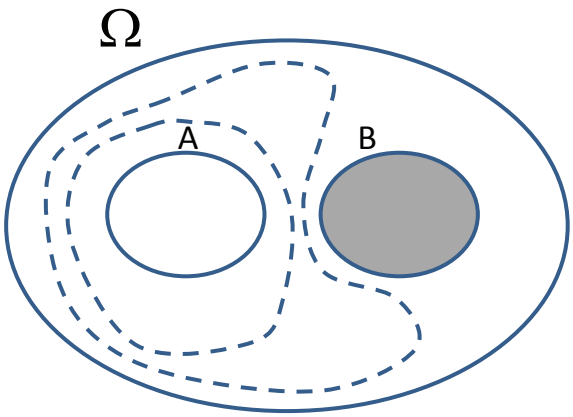

Figure 1: Two subsets of Ω (broken lines) containing A and not intersecting B. The set of all such subsets is denoted by $\varphi(A, B)$.

double-exponential complexity, making the approach inapplicable except for sets Ω with very small cardinality.

As an alternative, we propose to define mass functions and associated functions on a subset of 2^Θ that forms a lattice when equipped with the inclusion relation. The intuitive idea underlying our approach is the fact that, when expressing knowledge about a set-valued variable X, it is often convenient to specify sets of values that are *certainly* taken by X, and sets of values that are *certainly not* taken by X. This can be illustrated by the following example.

Example 1 Let X denote the languages spoken by John, defined on the (very large) set Ω of existing languages. If we know for sure that John can speak English and French (because he was brought up in the US and he stayed in France for a long time), and that he can speak neither Japanese nor Chinese (because he never traveled to Asia), then all subsets of Ω containing $A = \{English, French\}$ and not intersecting $B = \{Japanese, Chinese\}$ are possible values of X.

As shown by this example, some families of subsets of Ω or, equivalently, some subsets of $\Theta = 2^\Omega$ can be conveniently described by two subsets A and B of Ω such that $A \cap B = \emptyset$ (Figure 1).

More generally, let $\mathcal{Q}(\Omega) = \{(A, B) \in 2^\Omega \times 2^\Omega | A \cap B = \emptyset_\Omega\}$ be the set of ordered pairs

of disjoint subsets of Ω, where $\emptyset_\Omega$ denotes the empty set of Ω. For any $(A, B) \in \mathcal{Q}(\Omega)$, let $\varphi(A, B)$ denote the following subset of $\Theta = 2^\Omega$:

$$\varphi(A, B) = \{C \subseteq \Omega | C \supseteq A, C \cap B = \emptyset_\Omega\}. \tag{46}$$

$\varphi(A, B)$ is thus the subset of Θ composed of all subsets of Ω including A and non intersecting B. Equivalently, it is the set of all subsets of Ω that include A and are included in $\overline{B}$:

$$\varphi(A, B) = \{C \subseteq \Omega | A \subseteq C \subseteq \overline{B}\}. \tag{47}$$

It is thus the interval $[A, \overline{B}]$ in the lattice $(\Omega, \subseteq)$.

Let $\mathcal{C}(\Omega)$ denote the set of all subsets of Θ of the form $\varphi(A, B)$, completed by the empty set of Θ, noted $\emptyset_\Theta$:

$$\mathcal{C}(\Omega) = \{\varphi(A, B) | A \subseteq \Omega, B \subseteq \Omega, A \cap B = \emptyset_\Omega\} \cup \{\emptyset_\Theta\}.$$

$\mathcal{C}(\Omega)$ is thus a subset of 2^Θ. For a reason that will become evident later, we will also use $\varphi(\Omega, \Omega)$ as an alternative notation for $\emptyset_\Theta$. Function φ is thus a bijective mapping from $\mathcal{Q}^*(\Omega) = \mathcal{Q}(\Omega) \cup \{(\Omega, \Omega)\}$ to $\mathcal{C}(\Omega)$. The following proposition states that $\mathcal{C}(\Omega)$ is a closure system and, consequently, has a lattice structure.

Proposition 1 *$\mathcal{C}(\Omega)$ is a closure system of Θ, and*

$$\varphi(A, B) \cap \varphi(A', B') = \begin{cases} \varphi(A \cup A', B \cup B') & \text{if } (A \cup A') \cap (B \cup B') = \emptyset_\Omega \\ \emptyset_\Theta & \text{otherwise,} \end{cases}$$

for all (A, B) and (A', B') in $\mathcal{Q}^*(\Omega)$.

Proof: It is obvious that $\Theta = \varphi(\emptyset_\Omega, \emptyset_\Omega) \in \mathcal{C}(\Omega)$. Now,

$$\begin{aligned} \varphi(A, B) \cap \varphi(A', B') &= \{C \subseteq \Omega | C \supseteq A, C \supseteq A', C \cap B = \emptyset_\Omega, C \cap B' = \emptyset_\Omega\} \\ &= \{C \subseteq \Omega | C \supseteq (A \cup A'), C \cap (B \cup B') = \emptyset_\Omega\}. \end{aligned}$$

If $(A \cup A') \cap (B \cup B') = \emptyset_\Omega$ then $\varphi(A, B) \cap \varphi(A', B')$ is thus equal to $\varphi(A \cup A', B \cup B')$. Otherwise, no subset C of Ω can include $A \cup A'$ and have an empty intersection with $B \cup B'$; consequently, $\varphi(A, B) \cap \varphi(A', B') = \emptyset_\Theta$. □

As recalled in Section 3.1, any closure system endowed with the inclusion relation has a lattice structure with $\wedge = \cap$ and $\vee$ defined by (28). Here, the inclusion relation has the following simple expression using the $\varphi(A, B)$ representation:

$$\varphi(A, B) \subseteq \varphi(A', B') \Leftrightarrow A \supseteq A' \text{ and } B \supseteq B'. \tag{48}$$

The least element is $\bot = \varphi(\Omega, \Omega) = \emptyset_\Theta$. We note that (48) remains valid when $A = B = \Omega$, which explains the interest of the notation $\varphi(\Omega, \Omega) = \emptyset_\Theta$. The greatest element is $\top = \varphi(\emptyset_\Omega, \emptyset_\Omega) = \Theta$. The atoms are of the form $\varphi(A, \overline{A})$ for $A \subseteq \Omega$, and the co-atoms are of the form $\varphi(\{x\}, \emptyset_\Omega)$ or $\varphi(\emptyset_\Omega, \{x\})$ for $x \in \Omega$. We can see that the number of atoms is not equal to the number of co-atoms, which shows that $(\mathcal{C}, \subseteq)$ is not autodual. This lattice is also not complemented; consequently, it is not Boolean.

As a consequence of (48), it is easy to see that the meet operation $\vee$ is the following operation, hereafter denoted $\sqcup$:

$$\varphi(A, B) \sqcup \varphi(A', B') = \varphi(A \cap A', B \cap B').$$

It must be noted that $\sqcup$ is not identical to set union. The following proposition states the relation between these two operators.

Proposition 2 For all (A, B) and (A', B') in $\mathcal{Q}^*(\Omega)$,

$$\varphi(A, B) \cup \varphi(A', B') \ \subseteq \ \varphi(A, B) \sqcup \varphi(A', B').$$

Proof: For every C in $\varphi(A, B) \cup \varphi(A', B')$, we have

$$C \supseteq A \text{ and } C \supseteq A' \Rightarrow C \supseteq A \cap A' \tag{49}$$

and

$$C \cap B = \emptyset_\Omega \text{ and } C \cap B' = \emptyset_\Omega \Rightarrow C \cap (B \cap B') = \emptyset_\Omega, \tag{50}$$

hence $C \in \varphi(A \cap A', B \cap B')$. □

One can notice that the implications in (49) and (50) are strict. Consequently, $\varphi(A, B) \cup \varphi(A', B')$ is usually a strict subset of $\varphi(A, B) \sqcup \varphi(A', B')$. As the lattices $(\mathcal{C}(\Omega), \subseteq)$ and $(2^\Theta, \subseteq)$ do not have the same join operator, $(\mathcal{C}(\Omega), \subseteq)$ is not a sublattice of $(2^\Theta, \subseteq)$, although it is a subposet.

As noticed in [15], any ordered pair (A, B) of disjoint subsets of $\Omega = \{\omega_1, \ldots, \omega_K\}$ can be represented by a vector $(u_1, \ldots, u_K) \in \{-1, 0, 1\}^K$, with

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

Consequently, any $\varphi(A, B) \in \mathcal{C}(\Omega)$ such that $(A, B) \neq (\Omega, \Omega)$ can be represented in the same way. For $\varphi(\Omega, \Omega) = \emptyset_\Theta$, a special representation can be adopted, e.g., $(*, \ldots, *)$. 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 $\mathcal{C}(\Omega)$ is equal to $3^K + 1$.

4.2 Belief Functions on $\mathcal{C}(\Omega)$

The general theory recalled in Section 3.2 can be applied directly to the lattice $(\mathcal{C}(\Omega), \subseteq)$.

Let $m : \mathcal{C}(\Omega) \to [0, 1]$ be a mass function on $\mathcal{C}(\Omega)$. The notation $m(\varphi(A, B))$ will be simplified to $m(A, B)$. For this reason, m will be called a *two-place mass function*. We assume that

$$\sum_{(A,B) \in \mathcal{Q}^*(\Omega)} m(A, B) = 1.$$

The implicability, belief and commonality functions can be computed from m using the following formula:

$$b(A, B) = \sum_{\varphi(C,D) \subseteq \varphi(A,B)} m(C, D) = \sum_{C \supseteq A, D \supseteq B} m(C, D), \tag{51}$$

$$bel(A, B) = b(A, B) - m(\Omega, \Omega), \tag{52}$$

$$q(A, B) = \sum_{\varphi(C,D) \supseteq \varphi(A,B)} m(C, D) = \sum_{C \subseteq A, D \subseteq B} m(C, D), \tag{53}$$

where all pairs (A, B) and (C, D) are understood to belong to $\mathcal{Q}^*(\Omega)$ (the same convention will be adopted throughout this paper). The conjunctive sum operation in $\mathcal{C}(\Omega)$ is defined

as follows:

$$(m_1 \textcircled{\cap} m_2)(A,B) = \sum_{\varphi(C,D)\cap\varphi(E,F)=\varphi(A,B)} m_1(C,D)m_2(E,F) \tag{54}$$

$$= \begin{cases} \sum_{C\cup E=A, D\cup F=B} m_1(C,D)m_2(E,F) & \text{if } A\cap B=\emptyset_\Omega, \\ \sum_{(C\cup E)\cap(D\cup F)\neq\emptyset_\Omega} m_1(C,D)m_2(E,F) & \text{if } A=B=\Omega. \end{cases} \tag{55}$$

It can be computed using the commonality functions as:

$$q_{1\textcircled{\cap}2}(A,B) = q_1(A,B)\cdot q_2(A,B), \quad \forall (A,B)\in\mathcal{Q}^*(\Omega). \tag{56}$$

Similarly, the disjunctive sum can be computed as:

$$(m_1 \textcircled{\cup} m_2)(A,B) = \sum_{\varphi(C,D)\sqcup\varphi(E,F)=\varphi(A,B)} m_1(C,D)m_2(E,F) \tag{57}$$

$$= \sum_{C\cap E=A, D\cap F=B} m_1(C,D)m_2(E,F), \tag{58}$$

or using implicability functions:

$$b_{1\textcircled{\cup}2}(A,B) = b_1(A,B)\cdot b_2(A,B), \quad \forall (A,B)\in\mathcal{Q}^*(\Omega).$$

It is also possible to define a rule expressing a consensus among items of evidence, somehow in the same spirit as the Dubois-Prade rule recalled in Section 2.1. Assume that we learn from two sources that the value of X is in $\varphi(C,D)$ and in $\varphi(E,F)$, but that $\varphi(C,D)\cap\varphi(E,F)=\emptyset_\Theta$, i.e., $(C\cup E)\cap(D\cup F)\neq\emptyset_\Omega$, so that the two pieces of information are in conflict. It may still be safe to keep $(C\cup E)\setminus(D\cup F)$ as positive information, and $(D\cup F)\setminus(C\cup E)$ as negative information. Denoting by $\sqcap$ the following operation on $\mathcal{C}(\Omega)$:

$$\varphi(C,D)\sqcap\varphi(E,F) = \varphi\left((C\cup E)\setminus(D\cup F), (D\cup F)\setminus(C\cup E)\right),$$

we may define a new combination rule as

$$(m_1 \boxdot m_2)(A,B) = \sum_{\varphi(C,D)\sqcap\varphi(E,F)=\varphi(A,B)} m_1(C,D)m_2(E,F). \tag{59}$$

This rule will be referred to as the *consensus rule*. We note that operations $\sqcap$ and $\boxdot$ are not associative. However, they are quasi-associative, as it is possible to define a n-ary version of $\sqcap$ as:

$$\varphi(C_1,D_1)\sqcap\ldots\sqcap\varphi(C_n,D_n) = \varphi\left(\bigcup_{i=1}^n C_i \setminus \bigcup_{i=1}^n D_i, \bigcup_{i=1}^n D_i \setminus \bigcup_{i=1}^n C_i\right).$$

To compute functions m, w and v from q or b using (30), (33), (39) and (43), we need the expression of the Möbius function μ. It is given in the following proposition.

Proposition 3 The Möbius function on $(\mathcal{C}(\Omega), \subseteq)$ is given, for any (A, B) and (A', B') in $\mathcal{Q}^*(\Omega)$ by

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

Proof: The proof is similar to that of Theorem 2 in [15] with simple adaptations, due to the similarity between two-place belief functions on $\mathcal{C}(\Omega)$ and bi-capacities (see Section 5 below). □

This result allows us to compute m from b as:

$$m(A,B) = \sum_{C \supseteq A, D \supseteq B} (-1)^{|C\setminus A|+|D\setminus B|} b(C,D), \tag{60}$$

and from q as

$$m(A,B) = \sum_{C \subseteq A, D \subseteq B} (-1)^{|A\setminus C|+|B\setminus D|} q(C,D). \tag{61}$$

The conjunctive and disjunctive weight functions may be computed, respectively, as:

$$w(A,B) = \prod_{C \subseteq A, D \subseteq B} q(C,D)^{(-1)^{|A\setminus C|+|B\setminus D|+1}}, \quad \forall (A,B) \neq (\emptyset_\Omega, \emptyset_\Omega), \tag{62}$$

and

$$v(A,B) = \prod_{C \supseteq A, D \supseteq B} b(C,D)^{(-1)^{|C\setminus A|+|D\setminus B|+1}}, \quad \forall (A,B) \neq (\Omega, \Omega), \tag{63}$$

which makes it possible to use the cautious and bold rules in this context.

Example 2 Let X now denote the set of languages spoken by Bernard. Assume that we are 100 % sure that Bernard speaks no other language than Dutch (d), English (e) and French (f), so that we can restrict the domain of X to $\Omega = \{d, e, f\}$. Suppose that we have the following items of evidence:

1. Bernard is Belgian. Approximately 60 % of Belgians are Dutch-speaking, and 40 % of Belgians are French-speaking (we neglect here the small German-speaking community for simplicity). According to a recent survey, approximately 20 % of French-speaking Belgians declare to have good knowledge of Dutch, whereas around 50 % of members of the Dutch speaking community claim to have good knowledge of French.

2. Bernard studied three years in Canada, where most universities are English-speaking, and some are French speaking. Based on available evidence, we have a 0.7 degree of belief that Bernard studied in an English-speaking university, and a 0.15 degree of belief that he studied in a French-speaking one.

Each of these two items of evidence can be represented by a mass function on $\mathcal{C}(\Omega)$. According to the first item of evidence, approximately $(0.6 \times 0.5) \times 100 = 30\%$ of Belgians speak Dutch and no French, approximately $(0.4 \times 0.8) \times 100 = 32\%$ speak French and no Dutch, and the rest speak both languages. Knowing that Bernard belongs to this population (and nothing else), and assuming these figures to be accurate, this would lead to the following mass function:

$$m_1(\{d\}, \{f\}) = 0.3, \quad m_1(\{f\}, \{d\}) = 0.32, \quad m_1(\{f, d\}, \emptyset) = 0.38.$$

To account for inaccuracy of these figures, we may *discount* this mass function [26] by transferring a fraction of the mass (say, 10%) to the greatest element of $\mathcal{C}(\Omega)$, i.e., $\varphi(\emptyset, \emptyset)$. We thus have

$$m_1(\{d\}, \{f\}) = 0.3 \times 0.9 = 0.27, \quad m_1(\{f\}, \{d\}) = 0.32 \times 0.9 \approx 0.29,$$

$$m_1(\{f, d\}, \emptyset) = 0.38 \times 0.9 \approx 0.34, \quad m_1(\emptyset, \emptyset) = 0.1.$$

The second item of evidence can be represented by a mass function m_2 defined as:

$$m_2(\{e\}, \emptyset) = 0.7, \quad m_2(\{f\}, \emptyset) = 0.15, \quad m_2(\emptyset, \emptyset) = 0.15.$$

Assuming these two items of evidence to be distinct, they should be combined using the conjunctive sum operation $\textcircled{\cap}$. This may be achieved in two ways:

1. We may compute the intersection between each focal element of m_1 and each focal element of m_2 and apply formula (54). The computations may be presented as in Table 1.

2. Alternatively, we may compute the commonality functions q_1 and q_2 using (53), multiply them, and convert the result into a mass function using (61). The intermediate and final results are shown in Table 2.

Table 1: Computation of the conjunctive sum of m_1 and m_2 in Example 2. The columns and the lines correspond to the focal elements of m_1, and m_2, respectively. Each cell contains the intersection of a focal element of m_1 and a focal element of m_2. The mass of each focal element is indicated below it.

	$(\{d\},\{f\})$	$(\{f\},\{d\})$	$(\{f,d\},\emptyset)$	$(\emptyset,\emptyset)$
	0.27	0.29	0.34	0.1
$(\{e\},\emptyset)$	$(\{d,e\},f)$	$(\{e,f\},\{d\})$	$(\{e,f,d\},\emptyset)$	$(\{e\},\emptyset)$
0.7	0.7×0.27	0.7×0.29	0.7×0.34	0.7×0.1
$(\{f\},\emptyset)$	$\emptyset_\Theta$	$(\{f\},\{d\})$	$(\{f,d\},\emptyset)$	$(\{f\},\emptyset)$
0.15	0.15×0.27	0.15×0.29	0.15×0.34	0.15×0.1
$(\emptyset,\emptyset)$	$(\{d\},\{f\})$	$(\{f\},\{d\})$	$(\{f,d\},\emptyset)$	$(\emptyset,\emptyset)$
0.15	0.15×0.27	0.15×0.29	0.15×0.34	0.15×0.1

We may check that both approaches yield the same result. In particular, we can see that the empty set $\emptyset_\Theta$ receives a mass equal to $0.15 \times 0.27 = 0.0405$, which can be interpreted as a degree of conflict between m_1 and m_2. Using the consensus rule $\boxdot$ (59), the mass 0.15×0.27 would be transferred to

$$\varphi(\{f\},\emptyset) \sqcap \varphi(\{d\},\{f\}) = \varphi(\{d\},\emptyset),$$

resulting in a normal, conflict-free mass function.

Table 2 also shows the normal mass function computed using the normalized Dempster's rule $\oplus$, and Table 3 displays the intermediate steps and final results for computing the combinations of m_1 and m_2 using the unnormalized and normalized cautious rules.

Remark 2 We may remark here that the concept of two-place mass and belief functions defined here bears some similarity with bi-capacities introduced by Grabisch and Labreuche [15]. A bi-capacity as defined in [15] is an increasing function defined on the lattice $(\mathcal{Q}(\Omega),\sqsubseteq)$, where $\sqsubseteq$ is the partial ordering on $\mathcal{Q}(\Omega)$ defined by $(A,B) \sqsubseteq (C,D)$ if $A \subseteq B$ and $C \supseteq D$. In [15], Grabisch and Labreuche introduce various concepts related to bi-capacities, with application to cooperative game theory. In [19], they introduce the concept of bi-belief function, defined as a totally monotone bi-capacity from $\mathcal{Q}(\Omega)$ to $[0,1]$. They suggest an interpretation in terms of bipolar representation of uncertainty for the

Table 2: Computation of $m_1 \textcircled{\cap} m_2$ and $m_1 \oplus m_2$ in Example 2.

A	B	m_1	q_1	m_2	q_2	$q_1 \textcircled{\cap} 2$	$m_1 \textcircled{\cap} m_2$	$m_1 \oplus m_2$
$\{def\}$	$\{def\}$	0	1	0	1	1	0.0405	0
$\emptyset$	$\{def\}$	0	0.1	0	0.15	0.015	0	0
$\emptyset$	$\{de\}$	0	0.1	0	0.15	0.015	0	0
$\{f\}$	$\{de\}$	0	0.39	0	0.3	0.117	0	0
$\emptyset$	$\{df\}$	0	0.1	0	0.15	0.015	0	0
$\emptyset$	$\{d\}$	0	0.1	0	0.15	0.015	0	0
$\{f\}$	$\{d\}$	0.29	0.39	0	0.3	0.117	0.087	0.091
$\{e\}$	$\{df\}$	0	0.1	0	0.85	0.085	0	0
$\{e\}$	$\{d\}$	0	0.1	0	0.85	0.085	0	0
$\{ef\}$	$\{d\}$	0	0.39	0	1	0.39	0.203	0.212
$\emptyset$	$\{ef\}$	0	0.1	0	0.15	0.015	0	0
$\emptyset$	$\{e\}$	0	0.1	0	0.15	0.015	0	0
$\{f\}$	$\{e\}$	0	0.1	0	0.3	0.03	0	0
$\emptyset$	$\{f\}$	0	0.1	0	0.15	0.015	0	0
$\emptyset$	$\emptyset$	0.1	0.1	0.15	0.15	0.015	0.015	0.016
$\{f\}$	$\emptyset$	0	0.1	0.15	0.3	0.03	0.015	0.016
$\{e\}$	$\{f\}$	0	0.1	0	0.85	0.085	0	0
$\{e\}$	$\emptyset$	0	0.1	0.7	0.85	0.085	0.07	0.07
$\{ef\}$	$\emptyset$	0	0.1	0	1	0.1	0	0
$\{d\}$	$\{ef\}$	0	0.37	0	0.15	0.0555	0	0
$\{d\}$	$\{e\}$	0	0.1	0	0.15	0.015	0	0
$\{df\}$	$\{e\}$	0	0.44	0	0.3	0.132	0	0
$\{d\}$	$\{f\}$	0.27	0.37	0	0.15	0.0555	0.0405	0.0422
$\{d\}$	$\emptyset$	0	0.1	0	0.15	0.015	0	0
$\{df\}$	$\emptyset$	0.34	0.44	0	0.3	0.132	0.102	0.106
$\{de\}$	$\{f\}$	0	0.37	0	0.85	0.3145	0.189	0.197
$\{de\}$	$\emptyset$	0	0.1	0	0.85	0.085	0	0
$\{def\}$	$\emptyset$	0	0.44	0	1	0.44	0.238	0.248

Table 3: Computation of $m_1 \circledwedge m_2$ and $m_1 \circledwedge^* m_2$ in Example 2.

A	B	m_1	w_1	m_2	w_2	$w_{1\wedge 2}$	$m_1 \circledwedge m_2$	$m_1 \circledwedge^* m_2$
$\{def\}$	$\{def\}$	0	6.349	0	1	1	0.864	0
$\emptyset$	$\{def\}$	0	1	0	1	1	0	0
$\emptyset$	$\{de\}$	0	1	0	1	1	0	0
$\{f\}$	$\{de\}$	0	1	0	1	1	0	0
$\emptyset$	$\{df\}$	0	1	0	1	1	0	0
$\emptyset$	$\{d\}$	0	1	0	1	1	0	0
$\{f\}$	$\{d\}$	0.29	0.256	0	1	0.256	0.00806	0.0591
$\{e\}$	$\{df\}$	0	1	0	1	1	0	0
$\{e\}$	$\{d\}$	0	1	0	1	1	0	0
$\{ef\}$	$\{d\}$	0	1	0	1	1	0.0376	0.276
$\emptyset$	$\{ef\}$	0	1	0	1	1	0	0
$\emptyset$	$\{e\}$	0	1	0	1	1	0	0
$\{f\}$	$\{e\}$	0	1	0	1	1	0	0
$\emptyset$	$\{f\}$	0	1	0	1	1	0	0
$\emptyset$	$\emptyset$	0.1	10	0.15	6.67	719.6	0.00139	0.0102
$\{f\}$	$\emptyset$	0	1	0.15	0.5	0.5	0.00139	0.0102
$\{e\}$	$\{f\}$	0	1	0	1	1	0	0
$\{e\}$	$\emptyset$	0	1	0.7	0.176	0.176	0.00649	0.0477
$\{ef\}$	$\emptyset$	0	1	0	1.7	1	0.00649	0.0477
$\{d\}$	$\{ef\}$	0	1	0	1	1	0	0
$\{d\}$	$\{e\}$	0	1	0	1	1	0	0
$\{df\}$	$\{e\}$	0	1	0	1	1	0	0
$\{d\}$	$\{f\}$	0.27	0.270	0	1	0.270	0.00375	0.0275
$\{d\}$	$\emptyset$	0	1	0	1	1	0	0
$\{df\}$	$\emptyset$	0.34	0.227	0	1	0.227	0.00945	0.0694
$\{de\}$	$\{f\}$	0	1	0	1	1	0.0175	0.129
$\{de\}$	$\emptyset$	0	1	0	1	1	0	0
$\{def\}$	$\emptyset$	0	1	0	1	1	0.0441	0.324

Table 4: Commonalities of atoms according to $m_1 \oplus m_2$, $m_1 \boxdot m_2$ and $m_1 \circledwedge^* m_2$ in Example 2.

$(A, \overline{A})$	$q_{1\oplus 2}(A, \overline{A})$	$q_{1\boxdot 2}(A, \overline{A})$	$q_{1\wedge^* 2}(A, \overline{A})$
$(\emptyset, \{def\})$	0.0156	0.015	0.0102
$(\{f\}, \{de\})$	0.122	0.117	0.0796
$(\{e\}, \{df\})$	0.0889	0.085	0.0578
$(\{ef\}, \{d\})$	0.406	0.39	0.451
$(\{d\}, \{ef\})$	0.0578	0.096	0.0377
$(\{df\}, \{e\})$	0.138	0.173	0.0898
$(\{de\}, \{f\})$	0.328	0.355	0.214
$\mathbf{(\{def\}, \emptyset)}$	**0.459**	**0.481**	**0.509**

case of a single-valued variable. Bi-belief functions and two-place belief functions as introduced here are thus two distinct classes of belief functions built on different lattices, with different interpretations.

Remark 3 Another remark concerns decision making. As noted in the previous section, the lattice $(\mathcal{C}(\Omega), \subseteq)$ is not Boolean, so that the notion of pignistic probability cannot be defined in that lattice. In the classical setting, a common alternative to the rule of maximum pignistic probability for decision making is that of maximum plausibility: it consists in selecting the element of Ω with the greatest plausibility or, equivalent, with the greatest commonality (as these two functions coincide on singletons). In $\mathcal{C}(\Omega)$, we may propose as a reasonable decision rule to select the atom $\varphi(A, \overline{A})$ with the highest commonality. Table 4 shows the commonalities of the atoms computing from $m_1 \oplus m_2$, $m_1 \boxdot m_2$ and $m_1 \circledwedge^* m_2$ in Example 2. In that particular case, we can see that the three rules lead to the same conclusion, which is that Bernard speaks all three languages. The second most likely hypothesis is that Bernard speaks English and French, but no Dutch. However, it is clear that different combination rules may, in general, result in different decisions.

The following section will be devoted to a review of previous work on uncertainty representation for set-valued variables.

5 Relation to Previous Work

This section discusses the relation between the notions introduced above and related concepts or other formalisms already proposed for handling set-valued variables.

5.1 Disjunctive vs. Conjunctive Bodies of Evidence

Yager [32, 33] was among the first authors to emphasize the fundamental difference between single-valued and set-values variables, and to develop specific formalisms for reasoning with the latter. In [33], a distinction is made between *disjunctive* and *conjunctive* information using set-based representations. Given a variable X taking a single value in Ω, a statement "X is A" with $A \subseteq \Omega$ means that X takes *some* value in A, but we do not know which one. In contrast, if X is multiple-valued, the same statement is understood to mean that X takes *all* values in A (and possibly other values outside A). The corresponding piece of information is called "disjunctive" in the former case, and "conjunctive" in the latter. Yager then proceeds by observing that there is some kind of duality between disjunctive and conjunctive knowledge. For instance, the statement P_1 : "X is A" implies P_2 : "X is B" whenever $B \supseteq A$ in the disjunctive case, whereas P_2 can be deduced from P_1 whenever $B \subseteq A$ in the conjunctive case. If we know that P_1 and P_2 both hold, then we can deduce "X is $A \cap B$" in the disjunctive case, and "X is $A \cup B$" in the conjunctive case, etc.

Viewing mass functions as generalized sets, Dubois and Prade [8] remarked that the same distinction holds in the belief function framework. They pointed out that, when a mass function m represents a body of evidence pertaining to a set-valued variable (referred to as a conjunctive body of evidence), the commonality function q is more appropriate than b for representing degrees of belief, and the disjunctive sum (11) should be used for combining information conjunctively.

The formalism developed in Section 4 sheds new light on this duality between conjunctive and disjunctive knowledge. The conjunctive statement "X is A" corresponds to the proposition $\varphi(A, \emptyset)$. Let m be a mass function on $\mathcal{C}(\Omega)$ whose focal elements are all of the form $\varphi(B, \emptyset)$ for some $B \subseteq \Omega$. We can note $m'(A) = m(A, \emptyset)$ for all A. Using (51), we then have, for all $A \subseteq \Omega$:

$$b(A, \emptyset) = \sum_{B \supseteq A} m(B, \emptyset) = \sum_{B \supseteq A} m'(B) = q'(A),$$

where q' is the commonality function corresponding to m'. Conversely,

$$q(A, \emptyset) = \sum_{B \subseteq A} m(B, \emptyset) = \sum_{B \subseteq A} m'(B) = b'(A).$$

As a consequence, let m_1 and m_2 be two mass functions on $\mathcal{C}(\Omega)$ with focal elements of the form described above, and assume that we want to combine them conjunctively using (56). We get

$$q_{1 \textcircled{\cap} 2}(A, \emptyset) = q_1(A, \emptyset) q_2(A, \emptyset) = b'_1(A) b'_2(A) = b'_{1 \textcircled{\cup} 2}(A)$$

for all $A \subseteq \Omega$, which explains why the disjunctive sum *seems* to be used when combining conjunctive bodies of evidence in a conjunctive manner.

5.2 Random sets

Random sets are defined as random elements taking values as subsets of some space [20, 23]. In the finite case, a random set is thus defined by a probability function m on 2^Ω such that $\sum_{A \subseteq \Omega} m(A) = 1$, which is mathematically equivalent to a Dempster-Shafer mass function on Ω [24]. However, as noted by Smets [27], the semantics of random sets and (standard) belief functions are different, as random sets model random experiments with set-valued outcomes, whereas standard belief functions quantify beliefs regarding a variable taking a single, but unknown value.

In contrast, random sets are recovered as a special class of belief functions on set-valued variables introduced in this paper. Let m be a mass function on $\mathcal{C}(\Omega)$, and assume that the focal elements of m are atoms of $\mathcal{C}(\Omega)$, i.e., if they are of the form $(A, \overline{A})$. In that case, the function m' from 2^Ω to $[0, 1]$ such that $m'(A) = m(A, \overline{A})$ for all $A \subseteq \Omega$ is a random set. Random sets are thus equivalent to mass functions on $\mathcal{C}(\Omega)$ with atomic focal elements, just as probability distributions on Ω are equivalent to mass functions on 2^Ω with singleton focal elements.

5.3 Veristic Variables

In [34, 35], Yager develops a theory of *veristic* variables, which can be defined as fuzzy set-valued variables. Let X denote a fuzzy set-valued variable on a domain Ω, i.e., a variable taking a single value in the set I^Ω of fuzzy subsets of Ω. For any $x \in \Omega$ and any $A \in I^\Omega$, we denote by $A(x)$ the degree of membership of x in A. Let $A_0 \in I^\Omega$ denote

the unknown true value of X. Yager considers four ways of associating variable X with a fuzzy set $A \in I^\Omega$, using the following statements:

1. X *isv* A, meaning that $A \subseteq A_0$, where $\subseteq$ denotes standard fuzzy set inclusion, i.e., $A(x) \leq A_0(x)$ for all $x \in \Omega$;
2. X $isv(n)$ A, meaning that $A_0 \subseteq \overline{A}$, where $\overline{A}$ denotes the complement of A with membership function $\overline{A}(x) = 1 - A(x)$ for all $x \in \Omega$;
3. X $isv(c)$ A, meaning that $A_0 = A$;
4. X $isv(c, n)$ A, meaning that $A_0 = \overline{A}$.

In the above expressions, the copula *isv* has two parameters: c for closed (exclusive) and n for negative. We observe that the statement X $isv(c, n)$ A is equivalent to X $isv(c)$ $\overline{A}$. Consequently, we need only to consider the first three cases.

As remarked by Yager, each of these basic types of statements can be interpreted as specifying a set W of fuzzy subsets of Ω, i.e., a crisp subset of I^Ω. W contains the possible values of variable X. It is defined as follows for the three types of statements:

$$\begin{aligned} X \ isv \ A &\longrightarrow W = \{F \in I^\Omega \mid F(x) \geq A(x), \forall x \in \Omega\} \\ X \ isv(n) \ A &\longrightarrow W = \{F \in I^\Omega \mid F(x) \leq \overline{A}(x), \forall x \in \Omega\} \\ X \ isv(c) \ A &\longrightarrow W = \{A\} \end{aligned}$$

Yager also associated to W two functions from Ω to $[0, 1]$, called the *verity* and *rebuff* distributions, and defined as follows:

$$\mathrm{Ver}(x) = \min_{F \in W} F(x),$$

$$\mathrm{Rebuff}(x) = 1 - \max_{F \in W} F(x) = \min_{F \in W} 1 - F(x) = \min_{F \in W} \overline{F}(x).$$

$\mathrm{Ver}(x)$ is thus the minimal degree of membership of x to any possible value of X: it can be interpreted as the minimal support for x being one of the values taken by X. In contrast, $\mathrm{Rebuff}(x)$ can be interpreted as the minimal support for x *not* being one of the values taken by X. These two distributions have the following expressions for the three basic types of statements:

$$\begin{aligned} X \ isv \ A &\longrightarrow \mathrm{Ver}(x) = A(x), \quad \mathrm{Rebuff}(x) = 0 \\ X \ isv(n) \ A &\longrightarrow \mathrm{Ver}(x) = 0, \quad \mathrm{Rebuff}(x) = A(x) \\ X \ isv(c) \ A &\longrightarrow \mathrm{Ver}(x) = A(x), \quad \mathrm{Rebuff}(x) = 1 - A(x). \end{aligned}$$

Clearly, a major difference between Yager's approach and ours is the fact that Yager represents each piece of knowledge about X as a set of *fuzzy* subsets of Ω, whereas we use a set of *crisp* subsets of Ω. However, the kinds of statements considered by Yager as well as the associated verity and rebuff distributions have very close representations in our approach.

To begin with, let us provisionally assume that A is a crisp subset of Ω. Then, each of the three types of statements can be expressed by categorical mass functions on $\mathcal{C}(\Omega)$ as follows:

$$\begin{aligned} X\ isv\ A &\longrightarrow m(A,\emptyset)=1 \\ X\ isv(n)\ A &\longrightarrow m(\emptyset,A)=1 \\ X\ isv(c)\ A &\longrightarrow m(A,\overline{A})=1. \end{aligned}$$

It is easy to see that, in each of these three cases:

$$b(\{x\},\emptyset) = \text{Ver}(x) \tag{64}$$

$$b(\emptyset,\{x\}) = \text{Rebuff}(x) \tag{65}$$

for all $x \in \Omega$. The verity of x is thus the belief that x is one of the values taken by X, whereas the rebuff of x is the belief that x is not a value taken by X. This interpretation can be shown to remain true when A is a fuzzy subset of Ω. In that case, the function $x \rightarrow A(x)$ can be seen as a possibility distribution, which is known to be equivalent to a consonant mass function m' on Ω with focal elements $A_1 \subseteq \ldots \subseteq A_n$. The corresponding plausibility function pl' verifies

$$pl'(\{x\}) = \sum_{A_i \ni x} m'(A_i) = A(x), \quad \forall x \in \Omega.$$

For instance, let us consider the statement $X\ isv\ A$, and let us translate it as the following two-place mass function:

$$m(A_i,\emptyset) = m'(A_i), \quad i = 1,\ldots,n.$$

We have

$$b(\{x\},\emptyset) = \sum_{A_i \ni x} m(A_i,\emptyset) = \sum_{A_i \ni x} m'(A_i) = A(x) = \text{Ver}(x)$$

and

$$b(\emptyset,\{x\}) = 0 = \text{Rebuff}(x).$$

By handling the two other cases similarly, it can be verified that Equations (64) and (65) hold in all cases.

We may thus conclude that, although based on a slightly different interpretation, Yager's framework can be easily translated into the formalism of two-place belief functions, which is more general. However, this is only true at the *static* level, i.e., as long as we do not combine different pieces of information. For instance, as shown by Yager, the conjunctive combination of two statements X *isv* A and X *isv* B in the veristic framework results in a new statement X *isv* $A \cup B$, where $\cup$ denotes fuzzy set union. This is consistent with our approach only as long as A and B are crisp sets. If A and B are fuzzy, then translating the two statements as two-place mass functions and combining them using either the conjunctive sum or the cautious rule does not, in general, yield a consonant mass function corresponding to a veristic constraint on X. The two formalisms thus differ when combining statements involving fuzzy subsets.

5.4 Two-fold fuzzy sets

To complete this review of previous work on uncertainty representation for set-valued variables, we need to mention the representation of incomplete conjunctive information using a pair of fuzzy sets introduced in [9].

In this work, Dubois and Prade proposed to represent partial knowledge about a set-valued variable as a possibility distribution π on 2^Ω. This is equivalent to defining a fuzzy set of crisp subsets of Ω, which contrasts with Yager's approach who defines a crisp set W of fuzzy subsets of Ω. To make such a representation more easily tractable, Dubois and Prade then proposed to approximate π by a pair of fuzzy sets (A^-, A^+) as follows. Let A_i, $i \in I$ be the family of subsets of Ω such that $\pi(A_i) > 0$. Let

$$A^-(x) = 1 - \sup_{i: x \notin A_i} \pi(A_i)$$

and

$$A^+(x) = \sup_{i: x \in A_i} \pi(A_i).$$

The degree of membership of x to A^- is thus the extent to which is impossible to find an A_i not containing x, while $A^+(x)$ corresponds to the possibility of finding an A_i containing x. The pair (A^-, A^+), referred to as a *two-fold fuzzy set*, constitutes an approximation

of π in the sense that it is a simpler, but incomplete representation: several possibility distributions π correspond to the same two-fold fuzzy set. However, Dubois and Prade showed that the least specific possibility distribution π^* induced by a two-fold fuzzy set (A^-, A^+) can be expressed as $\pi^*(\emptyset) = 1 - \sup A^-$, $\pi^*(\Omega) = \inf A^+$, and

$$\pi^*(B) = \min \left[\inf_{x \in B} A^+(x), \inf_{x \notin B} (1 - A^-(x)) \right], \quad \forall B \in 2^\Omega \setminus \{\emptyset, \Omega\}.$$

To each two-fold fuzzy set (A^-, A^+) can thus be associated a fuzzy subset $\mathcal{A}$ of 2^Ω, with membership function equal to π^*.

We note that this approach has some similarity with ours, since it is based on the representation of a subset of 2^Ω by a pair of subsets of Ω. Actually, if A^- and A^+ are crisp, then the corresponding crisp subset $\mathcal{A}$ of 2^Ω is exactly equal to $\varphi(A^-, \overline{A^+})$. However, in the general case, the two-fold fuzzy set representation is based on a pair of possibility distributions, i.e., consonant belief functions on Ω, whereas our approach is based on a single two-place belief function on $\mathcal{C}(\Omega)$.

What can be seen as a limitation of the two-fold fuzzy set approach arises when combining information from several sources. Given two pairs (A^-, A^+) and (B^-, B^+) representing knowledge about two set-valued variables X and Y, Dubois and Prade showed that the knowledge of $X \cap Y$ can be represented by $(A^- \cap B^-, A^+ \cap B^+)$, while the knowledge of $X \cup Y$ can be represented by $(A^- \cup B^-, A^+ \cup B^+)$. Applications of this kind of reasoning to database query evaluation is discussed in [9]. However, a different and maybe more common problem in uncertain reasoning is the situation where we have two items of evidence about a single set-valued variable X, and we want to combine these two items of evidence. If (A^-, A^+) and (B^-, B^+) correspond, respectively, to fuzzy subsets $\mathcal{A}$ and $\mathcal{B}$ of 2^Ω, the result of the combination should ideally correspond to $\mathcal{A} \cap \mathcal{B}$ or to $\mathcal{A} \cup \mathcal{B}$, depending on the choice of a conjunctive or disjunctive combination mechanism. However, none of these two fuzzy subsets of 2^Ω generally admits a two-fold fuzzy set representation, which restricts the use of this formalism for reasoning with set-valued variables.

We have shown that the formalism of two-place belief functions introduced in this paper seems to compare favorably in terms of expressive power with existing formalisms for representing and reasoning with uncertain conjunctive information. In the next section, we will demonstrate the usefulness of this formalism for a certain category of classification problems.

6 Application to Multi-label Classification

In this section, we present an application of the framework developed in this paper to multi-label classification[2]. In this kind of problems, each object may belong simultaneously to several classes, contrary to standard single-label problems where objects belong to only one class [13, 38, 12, 36]. Multi-label classification tasks arise in many real-world problems. For instance, in image retrieval, each image may belong to several semantic classes such as beach and urban. In text categorization, each document may belong to several topics, etc. In such problems, the learning task consists in predicting 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 this paper may be used.

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 Ω of classes. 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. If several experts are consulted, some conflict will inevitably arise, which again will introduce some uncertainty in the labeling process.

The formalism developed in this paper 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 on $\mathcal{C}(\Omega)$. A less general, but arguably more workable option is to restrict m_i to be categorical, i.e., to have a single focal element $\varphi(A_i, B_i)$, with $A_i, B_i \subseteq \Omega$ and $A_i \cap B_i = \emptyset$. The set A_i is then the set of classes that *certainly apply* to example i, while B_i the set of classes that certainly *do not* apply. In a multiple expert setting, A_i might represent the set of classes indicated by all (or most) experts as relevant to describe instance i, while B_i would be the set of classes mentioned by none of the experts (or only a few of them). The usual situation of precise labeling is recovered in the special case where $B_i = \overline{A}_i$.

For instance, assume that instances are songs and classes are emotions generated by

[2] A preliminary version of the application described in this section was presented in [37].

these songs, as in the emotion dataset that will be used in Section 6.3 below. Upon hearing a song, an expert may decide that this song 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.

In [4, 39], we introduced a single-label k-nearest neighbor (NN) classifier based on Dempster-Shafer theory. This method will be briefly recalled in Section 6.1, and will be extended to multi-label classification tasks in Section 6.2. An experimental comparison with the multi-label k nearest neighbor (ML-kNN) method introduced in [38] using real-world data will then be presented in Section 6.3.

6.1 Single-label Evidential k-NN Classification

The evidential k-NN method introduced in [4] for single-label classification problems can be summarized as follows. Let $\mathcal{L} = \{(\mathbf{x}_1, A_1), \ldots, (\mathbf{x}_n, A_n)\}$ be a learning set of n instances, where $\mathbf{x}_i$ is a p-dimensional attribute vector describing instance i, and $A_i \subseteq \Omega = \{\omega_1, \ldots, \omega_K\}$ is a set of possible classes for instance i. We emphasize the fact that, in the context considered here, each instance i actually belongs to one and only one class, but this class is only known to lie somewhere in A_i.

Let $\mathbf{x}$ denote the feature vector for a new object with unknown class y. We want to guess the value of y based on evidence provided by the learning set $\mathcal{L}$. For that purpose, we consider the set $\Phi_k(\mathbf{x})$ of the k nearest neighbors of $\mathbf{x}$, according to some distance measure d (usually, the Euclidean one). Each learning instance $(\mathbf{x}_i, A_i)$ with $\mathbf{x}_i \in \Phi_k(\mathbf{x})$ can then be regarded as a piece of evidence regarding the unknown value of y, represented as the following simple mass function on Ω:

$$m_i(A_i) = \alpha \exp\left(-\gamma d(\mathbf{x}, \mathbf{x}_i)\right), \tag{66}$$

$$m_i(\Omega) = 1 - \alpha \exp\left(-\gamma d(\mathbf{x}, \mathbf{x}_i)\right), \tag{67}$$

with $0 < \alpha < 1$ and $\gamma > 0$. Parameter α is usually fixed at a value close to 1 such as $\alpha = 0.95$, whereas γ should depend on the scaling of distances and can be either fixed heuristically or optimized [39]. The evidence of the k NNs is then pooled using the

conjunctive sum:

$$m = \textcircled{\cap}_{i:\mathbf{x}_i \in \Phi(\mathbf{x})} m_i, \tag{68}$$

and the class with highest plausibility or pignistic probability is selected. As remarked in [7] and [6], this method can be easily extended to the case where each learning instance in $\mathcal{L}$ is labeled by a general mass function on Ω.

6.2 Multi-label Evidential k-NN Classification

Let us now come back to the multi-label classification problem, in which objects may belong *simultaneously* to several classes. Let $\mathcal{L} = \{(\mathbf{x}_1, A_1, B_1), \ldots, (\mathbf{x}_n, A_n, B_n)\}$ be the learning set, where $A_i \subseteq \Omega = \{\omega_1, \ldots, \omega_K\}$ denotes a set of classes that surely apply to instance i, and $B_i \subseteq \Omega$ a set of classes that surely do not apply to the same instance. If $Y_i \subseteq \Omega$ denotes the true label set of instance i, we thus only know that $Y_i \in \varphi(A_i, B_i)$.

As before, let $\Phi_k(\mathbf{x})$ denote the set of k nearest neighbors of a new instance described by feature vector $\mathbf{x}$, 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 simple two-valued mass function:

$$m_i(A_i, B_i) = \alpha \exp\left(-\gamma d(\mathbf{x}, \mathbf{x}_i)\right), \tag{69}$$

$$m_i(\emptyset, \emptyset) = 1 - \alpha \exp\left(-\gamma d(\mathbf{x}, \mathbf{x}_i)\right), \tag{70}$$

with, as before, $0 < \alpha < 1$ and $\gamma > 0$. These k mass functions are then combined using the conjunctive sum (68) as in the single-label case.

For decision making, different procedures can be used. The following simple and computationally efficient rule was implemented. Let $\widehat{Y}$ be the predicted label set for instance $\mathbf{x}$. To decide whether to include each class $\omega \in \Omega$ or not, we compute the degree of belief $bel(\{\omega\}, \emptyset)$ that the true label set Y contains ω, and the degree of belief $bel(\emptyset, \{\omega\})$ that it does not contain ω. We then define $\widehat{Y}$ as

$$\widehat{Y} = \{\omega \in \Omega \mid bel(\{\omega\}, \emptyset) \geq bel(\emptyset, \{\omega\})\}.$$

6.3 Experiments

To study the above procedure experimentally, three real datasets[3] were used:

[3]These datasets can be downloaded from `http://mlkd.csd.auth.gr/multilabel.html`.

- The *emotion dataset*, presented in [31], consist of 593 songs annotated by experts according to the emotions they generate. The emotions are: amazed-surprise, happy-pleased, relaxing-calm, quiet-still, sad-lonely and angry-fearful. Each emotion corresponds to a class. There are thus 6 classes, and each song was labeled as belonging to one or several classes. The average size of the label set for each song is 1.87 ± 0.67. Each song was also described by 8 rhythmic features and 64 timbre features, resulting in a total of 72 features. The data was split into a training set of 391 examples and a test set of 202 examples.

- The *yeast dataset* contains data regarding the gene functional classes of the yeast *Saccharomyces cerevisiae* [11, 25]. It describes 2417 genes each represented by 103 features. There are 14 possible classes and the average size of the label set for each gene is 4.24 ± 1.57. The data was split into a learning set of 1500 examples and a test set of 917 examples.

- The *scene dataset* consists of 2407 natural scene images, where a label set is manually assigned to each image. There are 6 classes and 294 attributes. The average cardinality is 1.074 ± 0.26 (only 7.35% of observations are labeled by more than one class). The data was split into a training set of 1211 examples and a test set of 1196 examples.

Each of these three datasets was constructed in such a way that each instance i is assigned a single set of labels Y_i. As explained above, this choice may be questioned since, at least for the emotion and scene datasets, there is no ground truth and the data have been labeled subjectively by a pool of experts. To assess the performances of our approach in learning from data with imprecise labels such as postulated in Section 6.2 above, we *randomly simulated an imperfect labeling process* by proceeding as follows.

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iK})$ be the vector of $\{-1, 1\}^K$ such that $y_{ik} = 1$ if $\omega_k \in Y_i$ and $y_{ik} = -1$ otherwise. For each instance i and each class ω_k, we generated a probability of error p_{ik} between 0 and 0.5 by drawing a random number from a beta distribution with parameters $a = b = 0.5$ (this is a bimodal distribution with modes at 0 and 1) and dividing it by two. We then changed y_{ik} to $-y_{ik}$ with probability p_{ik}, resulting in a noisy label

vector $\mathbf{y}'_i$. The imprecise label vector was finally defined as $\mathbf{y}''_i = (y''_{i1}, \ldots, y''_{iK})$ with

$$y''_{ik} = \begin{cases} y'_{ik} & \text{if } p_{ik} < 0.2, \\ 0 & \text{otherwise.} \end{cases}$$

As remarked in Section 4.1, such a vector of $\{-1, 0, 1\}^K$ encodes an ordered pair (A_i, B_i) of disjoint subsets of Ω such that $A_i = \{\omega_k \in \Omega \mid y''_{ik} = 1\}$ and $B_i = \{\omega_k \in \Omega \mid y''_{ik} = -1\}$.

The intuition behind the above model may be described as follows. Each number p_{ik} represents the probability that the membership of instance i to class ω_k will be wrongly assessed by the expert. This number may be turned into a degree of confidence c_i by the transformation $c_{ik} = 1 - 2p_{ik}$. We assume that these numbers can be provided by the expert, which allows us to label each instance i by a pair of sets (A_i, B_i). The set A_i then contains the classes ω_k that can be definitely assigned to instance i with a high degree of confidence ($c_{ik} \geq 0.6$), while B_i is the set of classes which are definitely *not* assigned to instance i. The remaining set $\Omega \setminus (A_i \cup B_i)$ contains those classes about which the expert is undecided ($c_{ik} < 0.6$).

Our method (hereafter referred to as EML-kNN) was applied to the three datasets, both with noisy labels $\mathbf{y}'_i$ and with imprecise labels (A_i, B_i). The features were normalized so as to have zero mean and unit variance. Parameters α and γ were fixed at 0.95 and 0.5, respectively, for all three datasets. We note that γ could easily be determined automatically by cross-validation. However, the results are not very sensitive to the value of γ, so that this parameter could be fixed manually.

As a reference method, we used the ML-kNN method introduced in [38], which was shown in [38] to have good performances as compared to most existing multi-label classification algorithms. It is also the closest to our method, as both methods are based on nearest neighbors. The ML-kNN algorithm was applied to noisy labels only, as it is not clear how imprecise labels could be handled using this method.

For evaluation, we used accuracy as a performance measure, defined as:

$$\text{Accuracy} = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i \cap \widehat{Y}_i|}{|Y_i \cup \widehat{Y}_i|},$$

where n is the number of test examples, Y_i is the true label set for examples i, and $\widehat{Y}_i$ is the predicted label set for the same example. This measure takes values between 0 and 1, with higher values corresponding to better performance.

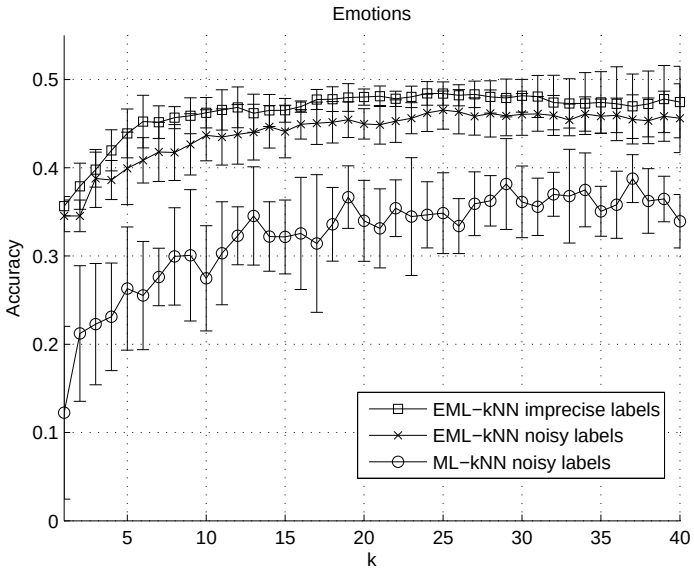

Figure 2: Mean accuracy (plus or minus one standard deviation) over 5 trials as a function of k for the emotions dataset with the following methods: EML-kNN with imprecise labels (A_i, B_i), EML-kNN with noisy labels and ML-kNN with noisy labels.

Figures 2 to 4 show the mean accuracy plus or minus one standard deviation over five generations of noisy and imprecise labels for the three datasets, with the following methods: EML-kNN with imprecise labels (A_i, B_i), EML-kNN with noisy labels and ML-kNN with noisy labels. The results are consistent over the three datasets: the EML-kNN method with noisy labels outperforms the ML-kNN trained using the same data, while the EML-kNN algorithm with imprecise labels clearly yields the best performances for the three problems.

These preliminary results demonstrate the ability of our approach to handle imprecise labels in multi-label classification tasks. More generally, they illustrate a practical situation where mass functions on a lattice $(\mathcal{C}(\Omega), \subseteq)$ are a natural model for expert knowledge and can be successfully exploited for uncertain reasoning with set-valued variables.

It should be noted, however, that these encouraging results are only a first step towards a comprehensive assessment of our approach in multi-label classification tasks. A more complete study would require more extensive comparisons with a wider range of algorithms and datasets, and more sophisticated schemes for tuning hyperparameters. Such a study

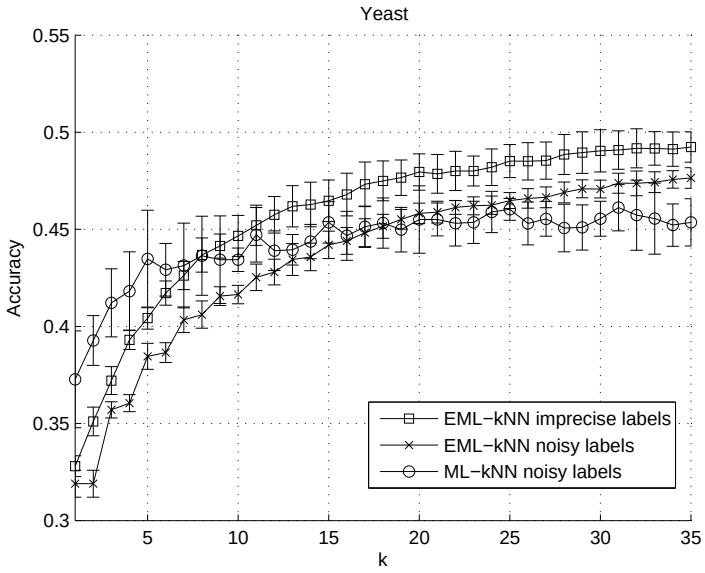

Figure 3: Mean accuracy (plus or minus one standard deviation) over 5 trials as a function of k for the yeast dataset with the following methods: EML-kNN with imprecise labels (A_i, B_i), EML-kNN with noisy labels and ML-kNN with noisy labels.

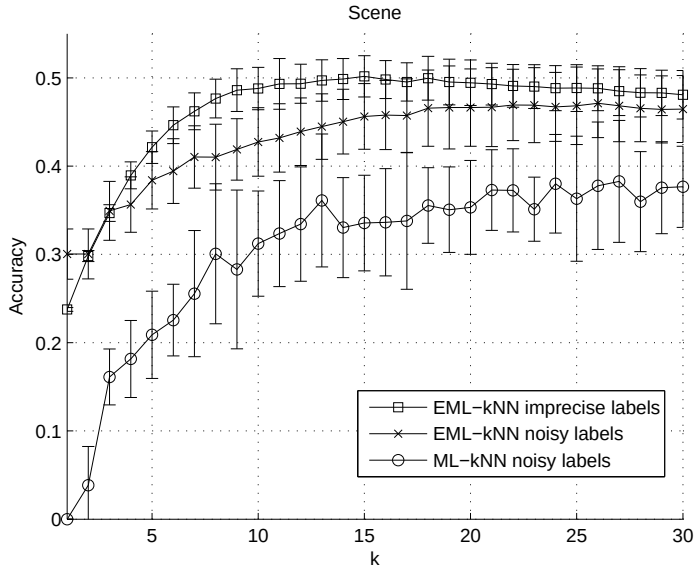

Figure 4: Mean accuracy (plus or minus one standard deviation) over 5 trials as a function of k for the scene dataset with the following methods: EML-kNN with imprecise labels (A_i, B_i), EML-kNN with noisy labels and ML-kNN with noisy labels.

goes beyond the scope of the present paper, and is left for future work.

7 Conclusion

We have presented a formalism for quantifying uncertainty on a set-valued variable X defined on a domain Ω in the belief function framework. This approach relies on the definition of a family $\mathcal{C}(\Omega)$ of subsets of 2^Ω that is closed under intersection and has a lattice structure. Each element C in this family is indexed by two subsets A and B, and is defined as the set of subsets of Ω containing A and not intersecting B. The number of such elements (including the empty set of 2^Ω) is equal to $3^K + 1$, where K is the size of Ω: it is thus much smaller than the size of 2^{2^Ω}, while being rich enough to express evidence about X in many realistic situations.

Using recent results about belief functions on general lattices reported in [14], we have shown that most notions from Dempster-Shafer theory can be defined on $\mathcal{C}(\Omega)$ with only a moderate increase in complexity as compared to the single-valued case, which contrasts with the double-exponential complexity encountered when working in 2^{2^Ω}. This formalism has been shown to be more general than previous attempts to apply the Dempster-Shafer framework to this problem. It has also been shown to be somewhat similar to, but arguably more general and flexible than other approaches introduced in the possibilistic framework.

Finally, our formalism has been applied to multi-label classification with imprecise labels, using an extension of the single-label evidential k nearest neighbor rule. Preliminary experimental results with real data and simulated uncertain labeling suggest that the proposed approach allows for the development of powerful classification procedures and can be applied to solve complex real-world problems. Further investigations into the belief function approach to multi-label classification, including extensive comparison with other methods, are currently under way and will be reported in future publications.

References

[1] S. Aguzzoli, B. Gerla, and V. Marra. De Finetti's no-Dutch-book criterion for Gödel logic. *Studia Logica*, 90(1), 2008.

[2] R. L. O. Cignoli, I. M. L. D'Ottaviano, and D. Mundici. *Algebraic foundations of many-valued reasoning*, volume 7 of *Trends in Logic-Studia Logica Library*. Kluwer Academic Publishers, Dordrecht, 2000.

[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] 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.

[5] T. Denœux. Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *Artificial Intelligence*, 172:234–264, 2008.

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

[7] T. Denœux and L. M. Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems*, 122(3):47–62, 2001.

[8] 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.

[9] D. Dubois and H. Prade. On incomplete conjunctive information. *Computers and Mathematics with Applications*, 15(10):797–810, 1988.

[10] D. Dubois and H. Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, 4:244–264, 1988.

[11] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 681–687. MIT Press, 2002.

[12] J. Fürnkranz, E. Hüllermeier, E. L. Mencia, and K. Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.

[13] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2004)*, pages 22–30, Sidney, Australia, 2004.

[14] M. Grabisch. Belief functions on lattices. *International Journal of Intelligent Systems*, 24:76–95, 2009.

[15] M. Grabisch and C. Labreuche. Bi-capacities – I: definition, Möbius transform and interaction. *Fuzzy Sets and Systems*, 151:211–236, 2005.

[16] T. Kroupa. Conditional probability on MV-algebras. *Fuzzy Sets and Systems*, 149(2):369–381, 2005.

[17] T. Kroupa. Representation and extension of states on MV-algebras. *Archive for Mathematical Logic*, 45(4):381–392, 2006.

[18] T. Kroupa. Belief functions on formulas in Lukasiewicz logic. In T. Kroupa and J. Vejnarová, editors, *8th Workshop on Uncertainty Processing (WUPES '09)*, Liblice, Czech Republic, 2009.

[19] C. Labreuche and M. Grabisch. Modeling positive and negative pieces of evidence in uncertainty. In T. D. Nielsen and N. L. Zhang, editors, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty (Proceedings of ECSQARU '03)*, pages 279–290, Aalborg, Denmark, 2003. Springer.

[20] G. Matheron. *Random Sets and Integral Geometry*. Wiley, New York, 1975.

[21] B. Monjardet. The presence of lattice theory in discrete problems of mathematical social sciences. Why. *Mathematical Social Sciences*, 46(2):103–144, 2003.

[22] D. Mundici. Averaging the truth-value in Lukasiewicz logic. *Studia Logica*, 55(1):113–127, 1995.

[23] H. Nguyen. *An Introduction to Random Sets*. Chapman and Hall/CRC Press, Boca Raton, Florida, 2006.

[24] H. T. Nguyen. On random sets and belief functions. *Journal of Mathematical Analysis and Applications*, 65:531–542, 1978.

[25] P. Pavlidis, J. Weston, J. Cai, and W. N. Grundy. Combining microarray expression data and phylogenetic profiles to learn functional categories using support vector machines. In *Proceedings of the Fifth Annual International Conference on Computational Biology*, pages 242–248, 2001.

[26] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.

[27] P. Smets. The Transferable Belief Model and random sets. *International Journal of Intelligent Systems*, 7:37–46, 1992.

[28] P. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

[29] P. Smets. The canonical decomposition of a weighted belief. In *Int. Joint Conf. on Artificial Intelligence*, pages 1896–1901, San Mateo, Ca, 1995. Morgan Kaufman.

[30] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.

[31] 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.

[32] R. R. Yager. On different classes of linguistic variables defined via fuzzy subsets. *Kybernetes*, 13:103–110, 1984.

[33] R. R. Yager. Set-based representations of conjunctive and disjunctive knowledge. *Information Sciences*, 41:1–22, 1987.

[34] R. R. Yager. Reasoning with conjunctive knowledge. *Fuzzy Sets and Systems*, 28:69–83, 1988.

[35] R. R. Yager. Veristic variables. *IEEE Transactions on Systems, Man and Cybernetics Part B*, 30(1):71–84, 2000.

[36] Z. Younes, F. Abdallah, and T. Denœux. Multi-label classification algorithm derived from k-nearest neighbor rule with label dependencies. In *6th European Signal Processing Conference (EUSIPCO '08)*, Lausanne, Switzerland, 2008.

[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] M.-L. Zhang and Z.-H. Zhou. ML-KNN: a lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[39] L. M. Zouhal and T. Denœux. An evidence-theoretic k-NN rule with parameter optimization. *IEEE Trans. on Systems, Man and Cybernetics C*, 28(2):263–271, 1998.