Dempster-Shafer reasoning in large partially ordered sets: Applications in Machine Learning

Thierry Denœux
Université de Technologie de Compiègne
Heudiasyc, UMR CNRS 6599
BP 20529 60205 Compiègne, France
Email: tdenoeux@hds.utc.fr

Marie-Hélène Masson
Université de Picardie Jules Verne
Heudiasyc, UMR CNRS 6599
BP 20529 60205 Compiègne, France
Email: mmasson@hds.utc.fr

***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, for instance, the power set of a finite set Ω, or the set of partitions of a finite set. Applications to multi-label classification and ensemble clustering are demonstrated.**

Keywords: Belief functions, lattice theory, set-valued variables, conjunctive knowledge, multi-label classification, ensemble clustering.

I. Introduction

The theory of belief functions originates from the pioneering work of Dempster [1], [2] and Shafer [3]. In the 1990's, the theory was further developed by Smets [4], [5], 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 [6], [7].

Although there has been some work on belief functions on continuous frames (see, e.g., [8], [9]), 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 $(2^\Omega, \subseteq)$ 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.

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, [10]). The complexity of manipulating and combining mass functions is then drastically reduced from 2^K to K^2. As we will show, 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 [11], belief functions can be defined on 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 will be shown, a lattice of intervals can be constructed, on which belief functions can be defined. This approach makes it possible to define belief functions on very large frames (such as the power set of a finite set Ω, or the set of partitions of a finite set) with manageable complexity.

The rest of this paper is organized as follows. The necessary background on belief functions defined on lattices will first be recalled in Section II. Our main idea will then be exposed in Section III. It will be applied to define belief functions on set-valued variables, with application to multi-label classification, in Section IV. The second example, presented in Section V, will concern belief functions on the set of partitions of a finite set, with application to ensemble clustering. Section VI will then conclude this paper.

II. Belief Functions on General Lattices

As shown by Grabisch [11], the theory of belief function can be defined not only on Boolean lattices, but on 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.

A. Lattices

A review of lattice theory can be found in [13]. The following presentation follows [11].

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 $\perp$) always exist. A strict partial ordering $<$ is defined from $\leq$ as $x < y$ if $x \leq y$ and $x \neq y$. 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 $\perp$. 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 $x \wedge x' = \perp$ 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 $\wedge = \cap$, $\vee = \cup$, $\perp = \emptyset$ and $\top = \Omega$.

A *closure system* on a set Θ is a family $\mathcal{C}$ of subsets of Θ containing Θ, and closed under intersection. As shown in [13], any closure system $(\mathcal{C}, \subseteq)$ is a lattice with $\wedge = \cap$ and $\vee = \sqcup$ defined by

$$A \sqcup B = \bigcap \{C \in \mathcal{C} | A \cup B \subseteq C\}, \quad \forall (A, B) \in \mathcal{C}^2. \tag{1}$$

B. 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:

$$f(x) = \sum_{y \leq x} m(y), \quad \forall x \in L. \tag{2}$$

Function m can be expressed as:

$$m(x) = \sum_{y \leq x} \mu(y, x) f(y), \tag{3}$$

where $\mu(x, y) : L^2 \to \mathbb{R}$ is the *Möbius function*, which is uniquely defined for each poset $(L, \leq)$. The *co-Möbius transform* of f is defined as:

$$q(x) = \sum_{y \geq x} m(y), \tag{4}$$

and m can be recovered from q as:

$$m(x) = \sum_{y \geq x} \mu(x, y) q(y). \tag{5}$$

Let us now assume that $(L, \leq)$ is a lattice. Following Grabisch [11], a function $b : L \to [0, 1]$ will be called an implicability function on L if $b(\top) = 1$, and its Möbius transform is non negative. The corresponding belief function bel can then be defined as:

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

Note that Grabisch [11] considered only normal belief functions, in which case $b = bel$. As shown in [11], any implicability function on $(L, \leq)$ is totally monotone. However, the converse does not hold in general: a totally monotone function may not have a non negative Möbius transform.

As shown in [11], most results of Dempster-Shafer theory can be transposed in the general lattice setting. For instance, the conjunctive sum between two mass functions m_1 and m_2 becomes:

$$(m_1 \circledcap m_2)(x) = \sum_{y \wedge z = x} m_1(y) m_2(z), \quad \forall x \in L, \tag{6}$$

and the following relation between commonality functions still holds:

$$q_{1 \circledcap 2}(x) = q_1(x) \cdot q_2(x), \quad \forall x \in L. \tag{7}$$

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

Similarly, the disjunctive rule is extended by substituting $\vee$ for $\cup$ in the usual definition.

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 can only be extended to autodual lattices [11].

III. 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 selecting as *events* only a strict subset of 2^Ω. As shown in Section II, the Dempster-Shafer calculus can be applied in this restricted set of events as long as it has a lattice structure. To be meaningful, the definition of events 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 $\leq$. The set of events is then defined as the set $\mathcal{I}$ of lattice intervals in $(\Omega, \leq)$. We will show that $(\mathcal{I}, \subseteq)$ is then itself a lattice, in which the Dempster-Shafer calculus can be applied.

This lattice $(\mathcal{I}, \subseteq)$ of intervals of a lattice $(\Omega, \leq)$ will first be introduced more precisely in Section III-A. The definition of belief functions on $(\mathcal{I}, \subseteq)$ will then be dealt with in Section III-B.

A. The Lattice $(\mathcal{I}, \subseteq)$

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 $\perp$. A subset I of Ω is a (lattice) interval if there exists 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 $[\perp, \top]$ and $\emptyset$ is the empty interval represented by $[a, b]$ for any a and

b such that $a \leq b$ does not hold. Let $\mathcal{I} \subseteq 2^{\Omega}$ be the set of intervals, including the empty set $\emptyset$:

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

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, $\mathcal{I}$ is a closure system, and $(\mathcal{I}, \subseteq)$ is a lattice, with least element $\emptyset$ and greatest element Ω. The meet operation is the intersection, and the join operation $\sqcup$ is defined by

$$[a,b] \sqcup [c,d] = [a \wedge c, b \vee d]. \tag{8}$$

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

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

B. Belief Functions on $(\mathcal{I}, \subseteq)$

Let m be a mass function from $\mathcal{I}$ to $[0,1]$. Implicability, belief and commonality functions can be defined on $(\mathcal{I}, \subseteq)$ as explained in Section II. Conversely, m can be recovered from b and q using (3) and (5), where the Möbius function μ depends on the lattice $(\mathcal{I}, \subseteq)$. As the cardinality of $\mathcal{I}$ 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.

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

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

Let b^* and q^* be the implicability and commonality functions associated to m^*. It is obvious that $b^*(I) = b(I)$ and $q^*(I) = q(I)$ for all $I \in \mathcal{I}$. Let m_1 and m_2 be two mass functions on $(\mathcal{I}, \subseteq)$, and let m_1^* and m_2^* be their "images" in $(2^{\Omega}, \subseteq)$. Because the meet operations are identical in $(\mathcal{I}, \subseteq)$ and $(2^{\Omega}, \subseteq)$, computing the conjunctive sum in any of these two lattices yields the same result, as we have

$$(m_1^* \textcircled{\cap} m_2^*)(A) = \begin{cases} (m_1 \textcircled{\cap} m_2)(A) & \text{if } A \in \mathcal{I}, \\ 0 & \text{otherwise.} \end{cases}$$

However, computing the disjunctive sum in $(2^{\Omega}, \subseteq)$ or $(\mathcal{I}, \subseteq)$ is not equivalent, because the join operation in $(\mathcal{I}, \subseteq)$, defined by (8), 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 image 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 an *outer approximation* [14], [15] of $m_1^* \textcircled{\cup} m_2^*$. When masses are assigned to intervals of the lattice $(\Omega, \leq)$, doing the calculations in $(\mathcal{I}, \subseteq)$ can thus be see an approximation of the calculations in $(2^{\Omega}, \subseteq)$, with a loss of information only when a disjunctive combination is performed.

IV. 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. The general framework will be presented in Section IV-A, and it will be applied to multi-label classification in Section IV-B.

A. 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* [14], [16]. 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 III [17].

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

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

where $\emptyset_{\Omega}$ denotes the empty sets of Ω (as opposed to the empty ste of Θ). Clearly, $\mathcal{I} \subseteq 2^{\Omega} = 2^{2^{\Theta}}$. 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 $\overline{B}$.

B. Multi-label Classification

In this section, we present an application of the framework developed in this paper to *multi-label classification* [18]–[20]. 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. For instance, in image retrieval, each image may belong to several semantic classes such as "beach" or "urban". 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 the previous section can be applied.

1) Training Data: 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. 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 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 in $(\mathcal{I}, \subseteq)$. 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. The usual situation of precise labeling is recovered in the special case where $A_i = B_i$.

2) Algorithm: The evidential k nearest neighbor rule introduced in [21] 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 $(\mathcal{I}, \subseteq)$:

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

where α and γ are two parameters such that $0 < \alpha < 1$. These k mass functions are then combined using the conjunctive sum.

For decision making, the following simple and computationally efficient rule can be used. Let $\widehat{Y}$ be the predicted label set for instance $\mathbf{x}$. To decide whether to include each class $\theta \in \Theta$ or not, we compute 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. We then define $\widehat{Y}$ as

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

3) Experiment: The *emotion dataset*[1], presented in [19], consist of 593 songs annotated by experts according to the emotions they generate. There are 6 classes, and each song was labeled as belonging to one or several classes. 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.

This dataset was initially constructed in such a way that each instance i is assigned a single set of labels Y_i. To assess the performances of our approach in learning from data with imprecise labels such as postulated in Section IV-B1 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 $\theta_k \in Y_i$ and $y_{ik} = -1$ otherwise. For each instance i and each class θ_k, we generated a probability of error $p_{ik} \in [0, 0.5]$ by drawing $2p_{ik}$ from a beta distribution with parameters a = b = 0.5, and we changed y_{ik} to $-y_{ik}$ with probability p_{ik}, resulting in a noisy label vector $\mathbf{y}'_i$. We then defined intervals $[A_i, B_i]$ such that $A_i = \{\theta_k \in \Theta \mid y'_{ik} = 1 \text{ and } p_{ik} < 0.2\}$ and $B_i = \{\theta_k \in \Theta \mid y'_{ik} = 1 \text{ or } p_{ik} \geq 0.2\}$.

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 θ_k will be wrongly assessed by the expert. We assume that these numbers can be provided by the expert as a way to describe the uncertainty of his/her assessments, which allows us to label each instance i by a pair of sets $[A_i, B_i]$.

Our method (hereafter referred to as EML-kNN) was applied 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. As a reference method, we used the ML-kNN method introduced in [18], which was shown to have good performances as compared to most existing multi-label classification algorithms. 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.

Figure 1 shows the mean accuracy plus or minus one standard deviation over five generations of noisy and imprecise labels, 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 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, which demonstrates the benefits of handling imprecise labels. This result was expected since our approach takes explicitly into account the additional information of error probabilities given by an expert.

V. Belief Functions on Partitions

Ensemble clustering methods [22], [23] 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 developed in this paper. 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

[1] This dataset can be downloaded from http://mlkd.csd.auth.gr/multilabel.html.

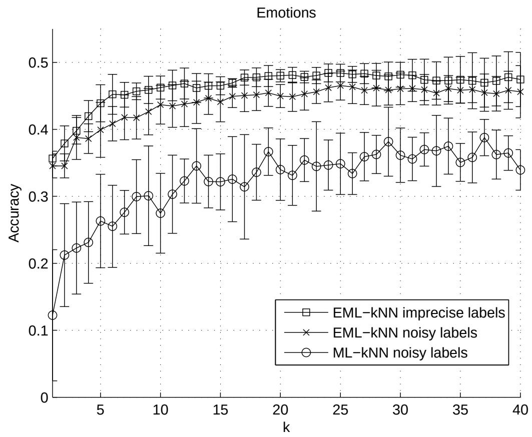

Figure 1. Mean accuracy (plus or minus one standard deviation) over 5 trials as a function of k for the emotions dataset with the following methods: EML-kNN with imprecise labels (A_i, B_i), EML-kNN with noisy labels and ML-kNN with noisy labels.

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 it will be explained below. Note that, due to space limitations, only the main principles will be given. More details may be found in [24], [25].

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

A. 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,...,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 $(x, y) \in E^2$, by:

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

The set of all partitions of E, denoted Ω, can be partially ordered using the following ordering relation: a partition p is said to be *finer* than a partition p' on the same set E ($p \preceq p'$) 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).

The set Ω endowed with the $\preceq$-order has a lattice structure [13]. 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 the ones that are coarser than p and p'. The equivalence relation $R_{p\vee p'}$ is given by the *transitive closure* of the maximum of R_p and $R_{p'}$. The least element of the lattice $\perp$ is the *finest* partition, denoted $p_0 = (1/2/.../n)$, in which each object is a cluster. The greatest element $\top$ of $(\Omega, \preceq)$ is the *coarsest* partition denoted $p_E = (123..n)$, in which all objects are put in the same cluster. 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.

A closed interval of Ω is defined as:

$$[\underline{p}, \overline{p}] = \{p \in \Omega \,|\, \underline{p} \preceq p \preceq \overline{p}\}. \tag{9}$$

It is a particular set of partitions, namely, the set of all partitions finer than $\overline{p}$ and coarser than $\underline{p}$.

B. Ensemble Clustering

1) Principle: We propose to use the following strategy for ensemble clustering:

1) Mass generation: Given r clusterers, build a collection of r mass functions m^1, m^2,..., m^r on the lattice of intervals; 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. An example will be given in Section V-B2.
2) Aggregation: Combine the r mass functions into a single one using the conjunctive sum. 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: Let p_{ij} denote the partition with $(n-1)$ clusters, in which the only objects which 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 put 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:

$$Bel_{ij} = bel([p_{ij}, p_E]) = \sum_{\underline{p}_k \succeq p_{ij}} m_k = \sum_{k=1}^{s} m_k \underline{R}_k(i,j). \tag{10}$$

Matrix $Bel = (Bel_{ij})$ can be considered as a new similarity matrix and can be in turn 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.

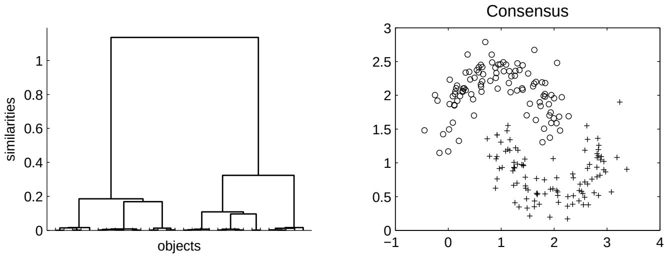

Figure 2. Half-rings data set. Ward's linkage computed from Bel and derived consensus.

2) Example: The data set used to illustrate the method is the half-ring data set inspired from [26]. 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).

Each hard partition p_l $(l = 1, 6)$ was characterized by a confidence degree $1-\alpha_l$, 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 of the clustering process, the following mass functions were defined:

$$\begin{cases} m^l([p_l, p_E]) = 1 - \alpha_l \\ m^l(\Omega) = \alpha_l. \end{cases} \tag{11}$$

The six mass functions (with two focal elements each) were then combined using the conjunctive rule of combination. A tree was computed from matrix Bel using Ward's linkage. This tree, represented in the left part of Figure 2, indicates a clear separation in two clusters. Cutting the tree to obtain two clusters yields the partition represented in the right part of Figure 2. We can see that the natural structure of the data is perfectly recovered.

VI. 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 if belief functions are defined over a subset of the power set 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, or the set of partitions of a set of objects. This approach opens the way to the application of Dempster-Shafer theory to computationally demanding Machine Learning tasks such as multi-label classification and ensemble clustering. Other potential applications of this framework include uncertain reasoning about rankings.

References

[1] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Annals of Mathematical Statistics*, vol. 38, pp. 325–339, 1967.

[2] ——, "Upper and lower probabilities generated by a random closed interval," *Annals of Mathematical Statistics*, vol. 39, no. 3, pp. 957–966, 1968.

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

[4] P. Smets, "The combination of evidence in the Transferable Belief Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 447–458, 1990.

[5] P. Smets and R. Kennes, "The Transferable Belief Model," *Artificial Intelligence*, vol. 66, pp. 191–243, 1994.

[6] G. Shafer, P. P. Shenoy, and K. Mellouli, "Propagating belief functions in qualitative Markov trees," *International Journal of Approximate Reasoning*, vol. 1, pp. 349–400, 1987.

[7] P. P. Shenoy, "Binary joint trees for computing marginals in the Shenoy-Shafer architecture," *International Journal of Approximate Reasoning*, vol. 17, pp. 239–263, 1997.

[8] L. Liu, "A theory of Gaussian belief functions," *International Journal of Approximate Reasoning*, vol. 14, pp. 95–126, 1996.

[9] P. Smets, "Belief functions on real numbers," *International Journal of Approximate Reasoning*, vol. 40, no. 3, pp. 181–223, 2005.

[10] T. Denœux, "Constructing belief functions from sample data using multinomial confidence regions," *International Journal of Approximate Reasoning*, vol. 42, no. 3, pp. 228–252, 2006.

[11] M. Grabisch, "Belief functions on lattices," *International Journal of Intelligent Systems*, vol. 24, pp. 76–95, 2009.

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

[13] B. Monjardet, "The presence of lattice theory in discrete problems of mathematical social sciences. Why," *Mathematical Social Sciences*, vol. 46, no. 2, pp. 103–144, 2003.

[14] D. Dubois and H. Prade, "A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets," *International Journal of General Systems*, vol. 12, no. 3, pp. 193–226, 1986.

[15] T. Denœux, "Inner and outer approximation of belief structures using a hierarchical clustering approach," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, no. 4, pp. 437–460, 2001.

[16] R. R. Yager, "Set-based representations of conjunctive and disjunctive knowledge," *Information Sciences*, vol. 41, pp. 1–22, 1987.

[17] T. Denoeux, Z. Younes, and F. Abdallah, "Representing uncertainty on set-valued variables using belief functions," *Submitted*, 2009.

[18] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: a lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[19] 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.

[20] 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)*, ser. LNAI, no. 5785. ashington, DC, USA: Springer-Verlag, 2009, pp. 297–308.

[21] T. Denœux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 25, no. 05, pp. 804–813, 1995.

[22] K.Hornik and F. Leisch., "Ensemble methods for cluster analysis," in *Adaptive Information Systems and Modelling in Economics and Management Science, Volume 5 of Interdisciplinary Studies in Economics and Management*, A. Taudes, Ed. Springer-Verlag, 2005, pp. 261–268.

[23] A. Fred and A. Lourenço, "Cluster ensemble methods: from single clusterings to combined solutions," *Studies in Computational Intelligence (SCI)*, vol. 126, pp. 3–30, 2008.

[24] M.-H. Masson and T. Denœux, "Belief functions and cluster ensembles," in *Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2009)*, C. Sossai and G. Chemello, Eds. Verona, Italy: Springer-Verlag, 2009, pp. 323–334.

[25] ——, "Ensemble clustering in the belief functions framework," *International Journal of Approximate Reasoning (submitted)*, 2010.

[26] A. Fred and A. Jain, "Data clustering using evidence accumulation," in *Proceedings of the 16th International Conference on Pattern Recognition*, Quebec, Canada, 2002, pp. 276–28.