Ensemble clustering in the belief functions framework

Marie-Hélène Masson[a,c], Thierry Denoeux[b,c]

[a] *Université de Picardie Jules Verne, IUT de l'Oise*
[b] *Université de Technologie de Compiègne*
[c] *Laboratoire Heudiasyc, UMR CNRS 6599*
BP 20529, 60205 Compiègne, France

Abstract

In this paper, belief functions, defined on the lattice of intervals partitions of a set of objects, are investigated as a suitable framework for combining multiple clusterings. We first show how to represent clustering results as masses of evidence allocated to sets of partitions. Then a consensus belief function is obtained using a suitable combination rule. Tools for synthesizing the results are also proposed. The approach is illustrated using synthetic and real data sets.

Key words: Clustering, ensemble clustering, belief functions, lattice of partitions, intervals of partitions

1. Introduction

Ensemble clustering methods [20, 16, 19] consist in combining multiple clustering solutions into a single one, called the *consensus*. The aim is to produce a more accurate and robust clustering of the data. Over several years a number of studies have been published on this topic[1]. The recent interest of the machine learning and artificial intelligence communities in ensemble techniques for clustering can be explained by the success of such ensemble techniques in a supervised setting. Moreover, some recent practical applications have shown the utility of such an approach in different contexts such as, e.g. in medical diagnosis [13, 18], gene expression microarray data clustering [28, 1], image segmentation [33]. Fundamental issues to be addressed when using ensemble clustering methods include: i) how to construct a set of individual solutions (or how to choose the base "clusterers"); and ii) how to combine the results of the ensemble into a single one.

This paper focuses on the second issue. Recent approaches to the problem of aggregating multiple clusterings include: voting schemes [8, 10], graph-based

[1] see, for example, the special issue of the Journal of Classification devoted to the "Comparison and Consensus of Classifications" published in 1986 [6]

approaches [37, 13], parameters estimation in a mixture of multinomial distributions [38], evidence accumulation clustering (EAC) [14, 15, 18]. This last approach is one of the most popular and will be shown to have some connections with the method proposed in this paper. In EAC, the collection of partitions is mapped onto a square co-association matrix where each cell (i, j) represents the fraction of times the pair of objects (x_i, x_j) has been assigned to the same cluster. This matrix is then considered as a similarity matrix which can in turn be clustered. A hierarchical clustering algorithm is often used for this purpose.

In this paper, which is an extension of [27], we address the problem of finding a consensus clustering as an information fusion problem in the general framework of uncertain reasoning. In fact, each clustering algorithm can be considered as a source, partially reliable, providing an opinion about the true, unknown, partition of the objects. The reliability of each source is assumed to be described by a confidence degree, either assessed by an external agent or evaluated using internal indices. An important point to consider is that, in some cases, the output of a clusterer does not provide evidence in favor of a single partition but supports naturally a set of possible hypotheses. This situation can occur in various circumstances, depending on the way the ensemble is generated, for example:

1. *Distributed clustering*: in that case, it is assumed that the whole data set is not available in a centralized location. Each clusterer has only a partial view (of small dimension) of the data and the combination is performed in a central location using only high level information such as cluster labels [37].
2. *Random subspaces, random hyperplanes*: for clustering high-dimensional data, some authors [12, 1] have proposed to generate their ensemble by randomly projecting the data on several low dimensional spaces. A similar approach is considered by Topchy et al. [39] who propose to combine multiple weak clusterings obtained by splitting the data by random hyperplanes.
3. *Complex shape clusters*: one way to generate a cluster ensemble is to split the data into a large number of small clusters [14]; different decompositions can be obtained by using different clustering algorithms or the same algorithm while varying a characteristic of the method (starting values, number of clusters, hyperparameter, order of presentation of the samples for on-line algorithms). The search for a partition compatible with all individual clusterings is a way to detect complex shape clusters.

In the first two cases, if a small number c of clusters is discovered in a given subspace, it seems natural to consider that the true number of clusters in the whole space is at least c. In other words, the information provided by each clusterer can be expressed as the set of all partitions that are at least as *fine* as the output of the clusterers. The third case correspond to the opposite situation: multiple clusterings of this type may be reconciled by assuming that the true unknown partition belongs to sets of partitions *coarser* than the individual ones.

There is thus a need to represent and manipulate information expressed as sets of partitions, possibly associated to confidence degrees. In this context, belief functions, a theory which has been successfully applied to many fusion and pattern recognition problems in recent years (sensors fusion, expert opinion pooling, classification), appear as a suitable framework for representing and combining the opinion of several clusterers. In this framework, a straightforward approach would be to consider, as the set of possible hypotheses (the frame of discernment), the set $\mathbb{P}$ of all possible partitions of the set to be clustered. Unfortunately, this approach requires algebraic manipulation of the elements of $2^{\mathbb{P}}$ and this can be intractable in the case where the number of partitions is high.

However, it is possible to work with a particular class of subsets of $2^{\mathbb{P}}$ (intervals of partitions), which will be shown to have a lattice structure. Some recent works ([17], [2]) have shown the possibility of defining and manipulating belief functions on any lattice. The use of a lattice structure allows us to dramatically limit the complexity when allocating belief masses to sets of partitions.

The rest of the paper is organized as follows. Section 2 gives the necessary background about lattices and belief functions. Section 3 focuses on partition lattices. Our approach is developed in Section 4. Some experimental results are presented in section 5. Finally, Section 6 concludes this paper.

2. Belief functions on general lattices

Lattices have recently attracted a great interest due to their high number of potential applications in computer science (databases, data mining, distributed computing, scheduling applications). This section begins with a short introduction to lattices. Section 2.2 gives the necessary background on belief functions which are classically defined on a Boolean lattice. Then the extension of belief functions to general lattices is presented.

2.1. Lattices

The following presentation follows [17]. Only the main definitions will be recalled. A more complete description on lattice theory can be found in [29].

A poset is a set P endowed with a partial order $\preceq$ (a reflexive, antisymmetric, transitive relation). A lattice is a poset $(P,\preceq)$ such that for any $x, y \in P$, their least upper bound $x \vee y$ and their greatest lower bound $x \wedge y$ exist. The element $x \vee y$, is called the *supremum* or *join* of x and y and $x \wedge y$ is called the *infimum* or *meet* of x and y. For finite lattices, there exist a greatest element, denoted $\top$, and a least element, denoted $\bot$. We say that *y covers x* if $x \preceq y$ and there is no z such that $x \preceq z \preceq y$. An element x 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 $\top$.

A lattice is *distributive* if $(x \vee y) \wedge z = (x \wedge z) \vee (y \wedge z)$ holds for all $x, y, z \in P$. A lattice $(P, \preceq)$ is said to be complemented if any $x \in P$ has a complement x' defined by $x \wedge x' = \bot$ and $x \vee x' = \top$. A lattice is *Boolean* if it is distributed and complemented. For any set Ω, the collection of all subsets of Ω, 2^Ω, ordered via subset inclusion, forms a lattice under the operations $\vee = \cup$ (set union) and $\wedge = \cap$ (set intersection).

2.2. Belief functions on a Boolean lattice

Dempster-Shafer theory of evidence (or belief functions theory) [34], like probability or possibility theories, is a theoretical framework for reasoning with partial and unreliable information. In this section, only the main concepts of this theory are recalled.

Let us consider a variable ω taking values in a finite set Ω called the frame of discernment. Partial knowledge regarding the actual value taken by ω can be represented by a *basic belief assignment* (bba) [34, 36], defined as a function m from 2^Ω to $[0, 1]$, verifying:

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

The subsets A of Ω such that $m(A) > 0$ are the *focal elements* of m. Each focal set A is a set of possible values for ω, and the quantity $m(A)$ can be interpreted as the measure of the belief that is committed exactly to $\omega \in A$ on the basis of a given evidential corpus. Complete ignorance corresponds to $m(\Omega) = 1$ (vacuous mass function), whereas perfect knowledge of the value of ω is represented by the allocation of the whole mass of belief to a unique singleton of Ω (m is then called a *certain* bba). A bba with nested focal elements is said to be *consonant*. A bba m is said to be of *simple support* if there exists $A \subset \Omega$ and $w \in [0, 1]$ such that $m(A) = 1 - w$ and $m(\Omega) = w$, all other masses being equal to zero. A bba m such that $m(\emptyset) = 0$ is said to be normal. The bba m can be equivalently represented by a credibility function bel, a plausibility function pl, and a commonality function q defined, respectively, by:

$$\mathrm{bel}(A) \triangleq \sum_{\emptyset \neq B \subseteq A} m(B) \quad \forall A \subseteq \Omega \ , \tag{2}$$

$$\mathrm{pl}(A) \triangleq \sum_{B \cap A \neq \emptyset} m(B) \quad \forall A \subseteq \Omega \ , \tag{3}$$

$$\mathrm{q}(A) \triangleq \sum_{B \supseteq A} m(B) \quad \forall A \subseteq \Omega \ . \tag{4}$$

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

$$\left\{ \begin{array}{l} m^\alpha(A) = (1 - \alpha) m(A) \quad \forall A \neq \Omega, \\ m^\alpha(\Omega) = (1 - \alpha) m(\Omega) + \alpha, \end{array} \right. \tag{5}$$

where $0 \leq \alpha \leq 1$ is the discount rate. This discount rate is related to the confidence held by an external agent in the reliability of the source [35]. It can be interpreted as the plausibility that the source is unreliable. When α is equal to 1, the vacuous mass function is obtained. When α=0, m remains unchanged.

Two bbas m^1 and m^2 induced by distinct items of evidence on Ω can be combined using the conjunctive rule of combination. The resulting mass function $m^1 \mathbin{\bigcirc\!\!\!\!\cap}\, m^2$ is defined by:

$$(m^1 \mathbin{\bigcirc\!\!\!\!\cap}\, m^2)(A) \triangleq \sum_{B \cap C = A} m^1(B) m^2(C) \quad \forall A \subseteq \Omega. \tag{6}$$

This conjunctive rule transfers the product of the masses $m^1(B)$ and $m^2(C)$ to the intersection of B and C. It can be expressed in a simple way using commonalities:

$$(\mathrm{q}^1 \mathbin{\bigcirc\!\!\!\!\cap}\, \mathrm{q}^2)(A) = \mathrm{q}^1(A)\mathrm{q}^2(A) \quad \forall A \subseteq \Omega, \tag{7}$$

where $\mathrm{q}^1(A)$ and $\mathrm{q}^2(A)$ denote, respectively, the commonalities associated to m^1 and m^2.

2.3. Extension to general lattices

As recalled in the previous section, belief functions are classically defined on the Boolean lattice 2^Ω. However, following initial investigations of Barthélemy [2], Grabisch [17] has shown that it is possible to extend these notions to the case where the underlying structure is no more a Boolean lattice, but any lattice: most results from Dempster-Shafer theory transfer to this general setting. Let $(P, \preceq)$ denote a lattice endowed with a $\vee$-meet and a $\wedge$-join operations. A basic belief assignment (bba) is defined as a mass function m from P to [0,1] verifying:

$$\sum_{x \in P} m(x) = 1. \tag{8}$$

The bba m can be equivalently represented by a credibility function bel, and a plausibility function q defined, respectively, by:

$$\mathrm{bel}(x) \triangleq \sum_{x' \preceq x} m(x') \quad \forall x \in P, \tag{9}$$

$$\mathrm{pl}(x) \triangleq \sum_{x \wedge x' \neq \perp} m(x') \quad \forall x \in P. \tag{10}$$

The conjunctive rule of combination is rewritten as:

$$(m^1 \mathbin{\bigcirc\!\!\!\!\cap}\, m^2)(x) = \sum_{x' \wedge x'' = x} m^1(x') m^2(x'') \quad \forall x \in P, \tag{11}$$

with the following relation between the commonalities:

$$(\mathrm{q}^1 \mathbin{\bigcirc\!\!\!\!\cap}\, \mathrm{q}^2)(x) = \mathrm{q}^1(x)\mathrm{q}^2(x) \quad \forall x \in P. \tag{12}$$

3. Lattices of partitions

In this section, we focus on a particular lattice structure on which belief functions may be defined: the lattice of partitions of a finite set. We first recall basic definitions about partitions and orders on partitions. The section ends with the presentation of another lattice derived from the previous one. This lattice, which seems particularly suitable in the context of consensus clustering, is composed of intervals of partitions. This makes it possible to manipulate sets of partitions with an acceptable level of complexity.

3.1. Partitions of a Finite Set

Let E denote a finite set of n objects. A partition p is a set of non empty subsets E_1,...,E_k of E such that:

1) the union of all elements of p, called clusters, is equal to E;
2) the elements of p are pairwise disjoint.

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:

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

Example. Let $E = \{1, 2, 3, 4, 5\}$. A partition p of E, composed of two clusters $\{1, 2, 3\}$ and $\{4, 5\}$ will be denoted as $p = (123/45)$. The associated equivalence relation is:

$$R_p = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

The set of all partitions of E, denoted by $\mathbb{P}_E$, 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 (or, equivalently p' is *coarser* than 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). In, this case, we write:

$$p \preceq p'.$$

This partial ordering can be alternatively defined using the equivalence relations associated to p and p':

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

The notation $\succeq$ will be used for the reverse relation defined by:

$$p' \succeq p \Leftrightarrow p \preceq p',$$

and that $\prec$, and $\succ$ will denote the restrictions of $\preceq$ and $\succeq$, respectively, to pairs of distinct elements:

$$p \prec p' \Leftrightarrow p \preceq p' \text{ and } p \neq p',$$

$$p \succ p' \Leftrightarrow p \succeq p' \text{ and } p \neq p'.$$

The *finest* partition ($\perp$) in the order $(\mathbb{P}_E, \preceq)$, denoted $p_0 = (1/2/.../n)$, is the partition in which each object is a cluster. The *coarsest* partition ($\top$) is $p_E = (123..n)$, in which all objects are put in the same cluster. Each partition precedes in this order every partition derived from it by aggregating two of its clusters. Similarly, each partition succeeds (*covers*) all partitions derived by subdividing one of its clusters in two clusters. The *atoms* of $(\mathbb{P}_E, \preceq)$ are the partitions preceded by p_0. There are $n(n-1)/2$ such partitions, each one having $(n-1)$ clusters with one and only one cluster composed of two objects. Atoms are associated with matrices R_p with only one off-diagonal entry equal to 1.

3.2. Lattice of Partitions

The set $\mathbb{P}_E$ endowed with the $\preceq$-order has a lattice structure [29]. *Meet* ($\wedge$) and *join* ($\vee$) operations can be defined as follows. The partition $p \wedge p'$, the *infimum* of p and p', is defined as the coarsest partition among all partitions finer than p and p'. The clusters of $p \wedge p'$ are obtained by considering pairwise intersections between clusters of p and p'. The equivalence relation $R_{p\wedge p'}$ is simply obtained by taking the minimum of R_p and $R_{p'}$. The partition $p \vee p'$, the *supremum* of p and 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 $R_{p'}$. Figures 1 and 2 show examples of partition lattices in the case where E is composed of three and four objects.

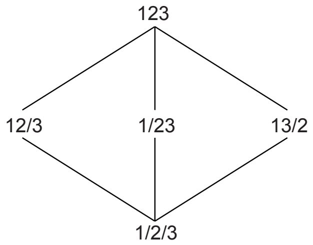

Figure 1: Lattice of partitions of a set of three elements.

3.3. Lattices of intervals of partitions

To enable the manipulation of sets of partitions, the previous framework has to be further extended in the following way. In $\mathbb{P}_E$, a closed interval of lattice elements is defined as:

$$[\underline{p}, \overline{p}] = \{p \in \mathbb{P}_E \,|\, \underline{p} \preceq p \preceq \overline{p}\}. \tag{13}$$

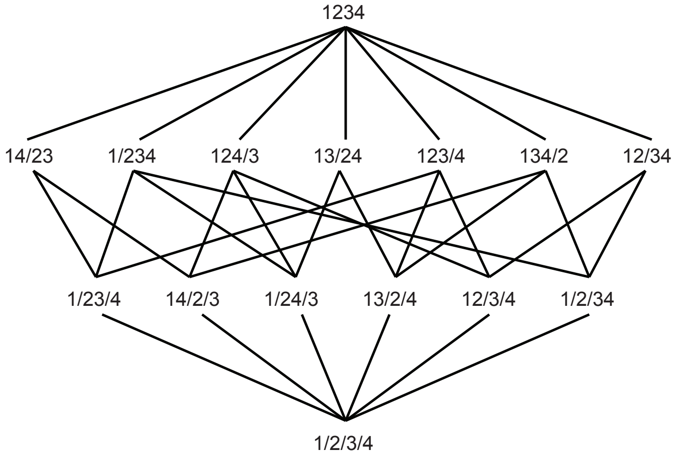

Figure 2: Lattice of partitions of a set of four elements.

An interval $[\underline{p}, \overline{p}]$ in $\mathbb{P}_E$ is thus a particular set of partitions, namely the set of all partitions finer than $\overline{p}$ and coarser than $\underline{p}$.

We consider now the set $\mathbb{I}_E$ of all intervals of $\mathbb{P}_E$, including the empty set of $\mathbb{P}_E$ (denoted by $\emptyset_{\mathbb{P}_E}$). The intersection of two intervals is also an interval:

$$[\underline{p}_1, \overline{p}_1] \cap [\underline{p}_2, \overline{p}_2] = \begin{cases} [\underline{p}_1 \vee \underline{p}_2, \overline{p}_1 \wedge \overline{p}_2], & \text{if } \underline{p}_1 \vee \underline{p}_2 \preceq \overline{p}_1 \wedge \overline{p}_2 \\ \emptyset_{\mathbb{P}_E} & \text{otherwise} \end{cases}. \tag{14}$$

So the set $\mathbb{I}_E$ is a closure system and, as shown by [29], is also a lattice, endowed with the inclusion relation:

$$[\underline{p}_1, \overline{p}_1] \subseteq [\underline{p}_2, \overline{p}_2] \Leftrightarrow \underline{p}_2 \preceq \underline{p}_1 \text{ and } \overline{p}_1 \preceq \overline{p}_2. \tag{15}$$

The *meet* operation is the intersection and the *join* of two elements $[\underline{p}_1, \overline{p}_1]$ and $[\underline{p}_2, \overline{p}_2]$ in $\mathbb{I}_E$ is defined as:

$$[\underline{p}_1, \overline{p}_1] \sqcup [\underline{p}_2, \overline{p}_2] = [\underline{p}_1 \wedge \underline{p}_2, \overline{p}_1 \vee \overline{p}_2]. \tag{16}$$

Note that the meet of two intervals corresponds exactly to the intersection of the corresponding sets of partitions, whereas the join of two intervals may be larger than the union of the sets of partitions.

In this lattice, the least element $\perp_{\mathbb{I}_E}$ is $\emptyset_{\mathbb{P}_E}$ and the greatest element $\top_{\mathbb{I}_E}$ is $\mathbb{P}_E$. The atoms of $\mathbb{I}_E$ are the singletons of $\mathbb{P}_E$. The co-atoms are of the form $[p_0, p]$ with p a co-atom of $(\mathbb{P}_E, \preceq)$ or $[p, p_E]$ with p an atom of $(\mathbb{P}_E, \preceq)$. An example of such a lattice, in the case where E is composed of three objects, is shown in Figure 3.

Within this framework, several kinds of imprecise knowledge about a partition can be expressed. For example, the intervals $[p_0, p]$ and $[p, p_E]$ represent the set of partitions finer and coarser, respectively, than a partition p. Suppose now that we know that the elements of a set $A \subseteq E$ are in the the same cluster. This external information is referred to as "must-link" constraints (see e.g. [40]) between the elements of A and can be translated as an interval of $\mathbb{I}_E$ as follows. Let p_A denote the partition in which the only elements which are clustered together are the elements of A:

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

Then the interval $[p_A, p_E]$ represents the set of all partitions in which the elements of A are clustered together. Note that "cannot link" constraints also used in constrained clustering, which specify that elements must not be clustered in the same class, can not be expressed in the proposed framework.

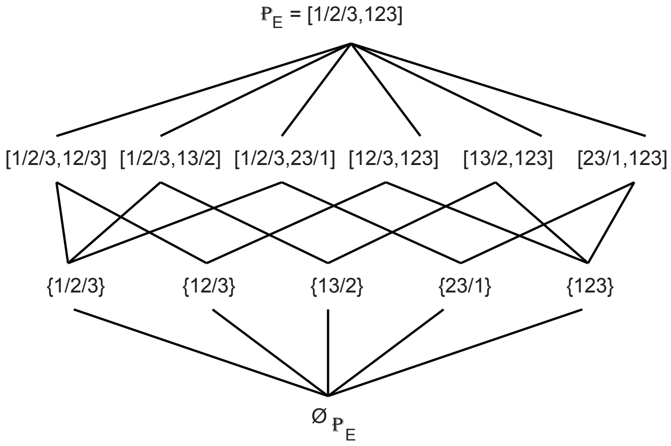

Figure 3: Lattice of intervals of partitions of a set composed of three elements.

4. Ensemble Clustering with partitions lattices

Belief functions defined on lattices of intervals of partitions, as introduced in the previous sections, offer a general framework for combining and synthesizing the results of several clustering algorithms. Note that an approach using the lattice of partitions (and not lattices of intervals of partitions, as it is proposed in this paper) for finding a consensus partition has been already suggested in the 80's by by Neumann and Norton [30]. Their approach works well for ensembles with little diversity. In case of strong conflicting opinions of the clusterers, their way of deriving a strict consensus, using meet and join operations on the lattice, can lead to very non informative results (the most trivial result is obtained when the meet is the set of singletons and the join is the partition with one cluster). This poor behavior is partly explained by the fact that, in this theoretical setting, the outputs of the clusterers are categorical opinions which are pooled together without the possibility of weighting them using mass functions.

We propose to use the following strategy for ensemble clustering:

1) Mass generation: Given r clusterers, build a collection of r bbas m^1,m^2,...,m^r on the lattice of intervals;
2) Aggregation: Combine the r bbas into a single one using the conjunctive rule of combination;
3) Synthesis: Provide a summary of the results.

These steps are discussed in detail below.

4.1. Basic belief assignment

A credal clustering ensemble is a collection of r bbas $\mathcal{C} = \{m^1, ..., m^r\}$. 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. Two representative examples are given below.

The simplest situation is encountered when a given clusterer l produces a single partition p_l of the data set (using for example the hard or fuzzy c-means algorithm). To account for the uncertainty of the clustering process, this categorical opinion can be transformed into a simple support mass function using the discounting operation (5). Let α_l denote the discounting factor of clusterer l (note that it is proposed in the experimental section to relate α_l to to an internal indice of validity of the partition). If the true unknown partition is considered to be at least as *fine* as p_l, the following mass allocation can be used:

$$\begin{cases} m_{l1} = m^l([p_0, p_l]) = 1 - \alpha_l \\ m_{l2} = m^l([p_0, p_E]) = \alpha_l. \end{cases} \tag{17}$$

On the contrary, if the true unknown partition is considered to be at least as *coarse* as the individual partition, then the following allocation may be considered:

$$\begin{cases} m_{l1} = m^l([p_l, p_E]) = 1 - \alpha_l \\ m_{l2} = m^l([p_0, p_E]) = \alpha_l. \end{cases} \tag{18}$$

Note that only two mass allocations are presented in this paper because they correspond to the practical cases mentioned in the introduction like complex shape clustering or random subspace clustering, but many other clustering methods could be described in the same framework. For instance, fuzzy equivalence relations, used for cluster analysis [4], or hierarchical clusterings, could be naturally represented by belief functions on the lattice of intervals of partitions.

The result of this mass allocation step is a collection of r bbas m^l which are defined, in the most general case, as a set of n_l focal elements $[\underline{p}^l_s, \overline{p}^l_s]$ with a mass $m^l_s = m^l([\underline{p}^l_s, \overline{p}^l_s])$ $(s = 1, ..., n_l)$:

$$m^l = \{([\underline{p}^l_s, \overline{p}^l_s], m^l_s), \quad s = 1, n_l\}$$

The equivalence relations associated to $\underline{p}_s^l$ and $\overline{p}_s^l$ will be denoted $\underline{R}_s^l$ and $\overline{R}_s^l$, respectively.

Two particular cases of ensembles will be of interest in the sequel. Type I ensembles will refer to ensembles composed exclusively of bbas m^l of the form:

$$m^l = \{([p_0, p_s^l], m_s^l), \quad s = 1, n_l\},$$

whereas type II ensembles will refer to ensembles composed of bbas m^l of the form:

$$m^l = \{([p_s^l, p_E], m_s^l), \quad s = 1, n_l\}.$$

4.2. Combination

Once the results provided by the r clusterers have been converted into r bbas, they can be aggregated into a single bba $m^* = m^1 \textcircled{\cap} m^2 \textcircled{\cap} ... \textcircled{\cap} m^r$ using the conjunctive rule of combination (11) with the meet operation defined by (14). The combination algorithm is summarized in appendix (algorithms 1 to 3).

The result of this combination is a bba m^*, i.e. a set of K intervals, associated with belief masses:

$$m^* = \{([\underline{p}_k^*, \overline{p}_k^*], m_k^*) \quad k = 1, K\}.$$

The associated equivalence relations of $\underline{p}_k^*$ and $\overline{p}_k^*$ will be denoted $\underline{R}_k^*$ and $\overline{R}_k^*$, respectively, in the sequel. Similarly, the credibility and plausibility functions related to m^* will be denoted bel* and pl*, respectively.

4.3. Synthesizing the results

The interpretation of the results is a difficult problem, since, depending on the number of clusterers in the ensemble, on their nature, on the conflict between them, and on the combination rule, a potentially high number K of focal elements may be found. If the number of focal elements in the combined bba is too high to be explored, a first way to proceed is to select only the focal elements associated with the highest masses. We propose also another approach which is explained below.

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 $(\mathbb{P}_E, \preceq)$). 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 two quantities, namely, the plausibility and the credibility of $[p_{ij}, p_E]$. They can be simply computed as follows:

$$
\begin{aligned}
Pl_{ij} = \mathrm{pl}^*([p_{ij}, p_E]) &= \sum_{[p_{ij},p_E]\cap[\underline{p}^*_k,\overline{p}^*_k]\neq\emptyset_{\mathbb{P}_E}} m^*_k && (19)\\
&= \sum_{\overline{p}^*_k \succeq p_{ij}} m^*_k && (20)\\
&= \sum_{k=1}^{K} m^*_k \overline{R}^*_k(i,j), && (21)
\end{aligned}
$$

$$
\begin{aligned}
Bel_{ij} = \mathrm{bel}^*([p_{ij}, p_E]) &= \sum_{[\underline{p}^*_k,\overline{p}^*_k]\subseteq[p_{ij},p_E]} m^*_k && (22)\\
&= \sum_{\underline{p}^*_k \succeq p_{ij}} m^*_k && (23)\\
&= \sum_{k=1}^{K} m^*_k \underline{R}^*_k(i,j). && (24)
\end{aligned}
$$

Note that, in case of a type I ensemble, the meet of any two focal elements $[p_0, p^l_s]$ and $[p_0, p^{l'}_{s'}]$ is equal to $[p_0, p^l_s \wedge p^{l'}_{s'}]$ so the combined bba m^* is such that:

$$\underline{p}^*_k = p_0 \quad \forall k = 1, K.$$

In that case, one has:

$$Bel_{ij} = \mathrm{bel}^*([p_{ij}, p_E]) = \sum_{k=1}^{K} m^*_k R_0(i,j) = 0, \quad \forall i \neq j,$$

where R_0 denotes the equivalence relation associated to p_0. In case of a type II ensemble, the meet of any two focal elements $[p^l_s, p_E]$ and $[p^{l'}_{s'}, p_E]$ is equal to $[p^l_s \vee p^{l'}_{s'}; p_E]$ so the combined bba m^* is such that:

$$\overline{p}^*_k = p_E \quad \forall k = 1, K.$$

In that case, one has:

$$Pl_{ij} = \mathrm{pl}^*([p_{ij}, p_E]) = \sum_{k=1}^{K} m^*_k R_E(i,j) = 1, \quad \forall i \neq j,$$

where R_E denotes the equivalence relation associated to p_E.

Matrices $Pl = (Pl_{ij})$ and $Bel = (Bel_{ij})$ can be considered as new similarity matrices and can be in turn clustered using, for instance, a hierarchical clustering algorithm. If a partition is needed, the classification tree can be cut at a specified level or so as to insure a user-defined number of clusters.

4.4. Special case of type I ensembles

In case of type I ensembles, the computation of Pl_{ij} can be further simplified. In fact, the particular shape of the focal elements of m^* allows us to write Pl_{ij} as:

$$Pl_{ij} = \mathrm{pl}^*([p_{ij}, p_E]) = q^*([p_0, p_{ij}]), \tag{25}$$

where q^* denotes the commonality function associated to m^*. As it is recalled in Section 2.3 in Eq. (12), q^* can be expressed as the product of the commonalities q^1,$\cdots$,q^r associated to m^1,$\cdots$,m^r, respectively:

$$Pl_{ij} = \prod_{l=1}^{r} q^l([p_0, p_{ij}]). \tag{26}$$

We have:

$$q^l([p_0; p_{ij}]) = \sum_{p_s^l \succeq p_{ij}} m_s^l = \sum_{s=1}^{n_l} m_s^l R_s^l(i,j). \tag{27}$$

Thus:

$$\ln Pl_{ij} = \sum_{l=1}^{r} \ln \left(\sum_{s=1}^{n_l} m_s^l R_s^l(i,j) \right). \tag{28}$$

Equation (28) shows that, in case of type I ensembles, it is not necessary to compute the result of the conjunctive rule of combination, because each Pl_{ij} can be simply computed from the initial bbas of the ensemble.

4.5. Link with the EAC approach

We assume in this section that the ensemble is of type I and that all bbas m^l are simple mass functions obtained by discounting categorical opinions given by the clusterers. We further assume that the discount factor is equal to α for all clusterers, so that:

$$m^l = \{([p_0, p^l], 1-\alpha), ([p_0, p_E], \alpha)\}.$$

In that case, the following equations hold:

$$q^l([p_0, p_{ij}]) = \begin{cases} 1 & \text{if} \quad p^l \succeq p_{ij} \\ 1-\alpha & \text{if} \quad p^l \prec p_{ij} \end{cases}, \tag{29}$$

and

$$Pl_{ij} = \prod_{\{l / p^l \succeq p_{ij}\}} 1 \prod_{\{l' / p^{l'} \prec p_{ij}\}} (1-\alpha) = (1-\alpha)^{\left(r - \sum_{l=1}^{r} R^l(i,j) \right)},$$

or, equivalently:

$$\ln Pl_{ij} = \left(r - \sum_{l=1}^{r} R^l(i,j) \right) \ln(1-\alpha). \tag{30}$$

One can see that $\ln Pl_{ij}$ is an increasing function of the fraction of times where i and j have been assigned to the same cluster by the individual clusterers. Consequently, clustering matrices Pl or $\frac{1}{r}\sum_l R^l$ using the single or complete linkage hierarchical clustering algorithm will yield the same results. This particular case is thus equivalent to the evidence accumulation approach. A weighted version of EAC, that weights differently the partitions in the co-association matrix, has also been proposed in [9]. It turns out that it is equivalent to our approach when each clusterer has its own discount rate α_l, as it can easily be shown that:

$$\ln Pl_{ij} = \sum_{l=1}^{r} \ln(1-\alpha_l)\left(1 - R^l(i,j)\right). \tag{31}$$

4.6. Toy example

Let $E = \{1,2,3,4,5,6,7\}$ be a set of 7 objects. We assume that two clustering algorithms have produced partitions $p_1 = (123/45/67)$ and $p_2 = (12/345/67)$. As it can be seen, the partitions disagree on the third element which is clustered with $\{1,2\}$ in p_1 and with $\{4,5\}$ in p_2. As proposed in Section 4, we construct two simple mass functions by discounting each clusterer l by a factor α_l. In a first situation, we consider that we have an equal confidence in the two clusterers, so we fix $\alpha_1 = \alpha_2 = 0.1$. Moreover, we assume that the unknown partition is finer than p_1 and p_2 (type I assignment). We have:

$$m^1([p_0,p_1]) = m^2([p_0,p_2]) = 0.9, \quad m^1([p_0,p_E]) = m^2([p_0,p_E]) = 0.1.$$

Applying Dempster's rule of combination (11) leads to the following combined bba $m^* = m^1 \textcircled{\cap} m^2$:

Focal elements	mass m^*	bel*
$[p_0, p_1 \wedge p_2]$	0.81	0.81
$[p_0, p_1]$	0.09	0.90
$[p_0, p_2]$	0.09	0.90
$[p_0, p_E]$	0.01	1

with $p_1 \wedge p_2 = (12/3/45/67)$.

A type II assignment, corresponding to the hypothesis that the true partition is coarser than the individual ones leads to the following combined mass $m*$:

Focal elements	mass m^*	bel*
$[p_1 \vee p_2, p_E]$	0.81	0.81
$[p_1, p_E]$	0.09	0.90
$[p_2, p_E]$	0.09	0.90
$[p_0, p_E]$	0.01	1

with $p_1 \vee p_2 = (12345/67)$. The matrices Pl and Bel computed from m^* are represented in the upper part of Figure 4. Logically, the type I assignment leads to a partition of the set into 4 clusters, whereas the type II assignment shows a structure into 2 clusters.

Suppose now that the confidence in the second clusterer is less than in the first one. Two situations are considered: we first fix $\alpha_1 = 0.1$ and $\alpha_2 = 0.5$ and then $\alpha_1 = 0.1$ and $\alpha_2 = 0.9$. The corresponding matrices Pl and Bel are represented in the middle and lower parts of Figure 4. As expected, the more the opinion of the second clusterer is discounted, the closer the combined partitions, whatever their type, to the partition given by the first clusterer.

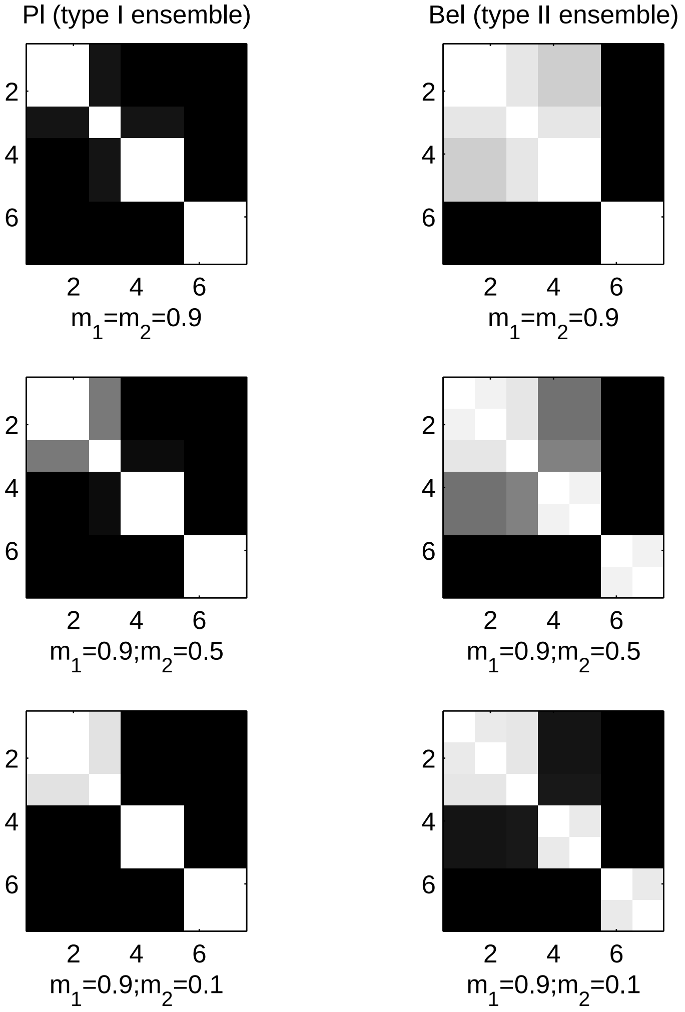

Figure 4: Plausibility and belief matrices for the toy example (white=1, black=0).

5. Some examples of applications

5.1. Preamble: discount rates

When a mass allocation such as (17) or (18) is used, fixing the discount rates α_l is a major issue. Each discounting factor has to reflect the confidence held in the clusterer. A way to automatically determine these factors is to relate them to a cluster validity index. The notion of cluster validation refers to concepts and methods for the quantitative evaluation of the output of a clustering algorithm.

Various cluster validity indices have been proposed to measure the quality of clustering results. They can be broadly divided into *external* and *internal* indices. External validity indices assess the agreement between a clustering solution and a predefined reference clustering. One of the most popular is the Rand Index [31], which determines the similarity between two partitions as a function of positive and negative agreements in pairwise cluster assignments. The Adjusted Rand index [21] (AR) introduces a normalization in order to yield values close to zero for random partitions. It is computed as follows. Let

- p_1 and p_2 be two partitions of a set E;
- a denote the number of pairs of elements in E that are in the same cluster in p_1 and in the same cluster in p_2;
- b denote the number of pairs of elements in E that are in different clusters in p_1 and in different clusters in p_2;
- c denote the number of pairs of elements in E that are in the same cluster in p_1 and in different clusters in p_2;
- d denote the number of pairs of elements in E that are in different clusters in p_1 and in the same cluster in p_1.

The Adjusted Rand index, AR, is defined as:

$$ARI(p_1, p_2) = \frac{2(ab - cd)}{(a+d)(d+b) + (a+c)(c+b)}. \tag{32}$$

This index belongs to [-1,1] and is close to 1 when the two partitions are in good agreement. In real-life unsupervised tasks, there is no reference clustering, so that such criterion is not directly applicable for deriving a discounting factor. However, the Adjusted Rand index will be used in the experiments to evaluate the quality of the final clustering with respect to known labels (when they are known) and also in the definition of an internal validity criterion as explained below.

Internal validation methods compare different solutions based on the goodness of fit between each clustering and the data by using combined notions of compactness, separation, and connectedness of the clusters. Examples of such indices are Dunn's index [11], the Davies-Bouldin index [5], and the Silhouette score [32]. However, these indices can only make comparisons between clusterings generated using the same metric. When, for example, a situation of distributed clustering is considered, these indices are not applicable. More recent approaches to the problem of cluster validation suggest to use the stability of a partitioning as an internal validation measure. In particular, we adopt the approach of [24, 3] based on the idea that a clustering algorithm should produce consistent results when applied to data sampled from the same source. The general principle is to repeatedly draw two overlapping subsamples of the same dataset. Each subsample is clustered individually, and the two resulting partitions are compared by applying an external validation index (we use the AR index) to the partial partitions obtained for the overlapping shared set of points. The average of the AR index over several repetitions is retained as a stability index of the clusterer (note that the AR index has been truncated to zero considering that negative AR values were associated to bad agreements between partitions). The detailed algorithm for computing the stability index is given in appendix (algorithm 4). This stability index has been used in all the experiments reported below (with a number of bootstrap samples $n_{boot} = 30$ and a sampling ration $S = 0.9$), by defining the discount rate of each clusterer as one minus the stability index.

5.2. Distributed clustering: first experiment

In a distributed computing environment, the data set is spread into a number of different sites. In that case, each clusterer has access to a limited number of features and the distributed computing entities share only higher level information describing the structure of the data such as cluster labels. The problem is to find a clustering compatible with what could be found if the whole set of features was considered. To illustrate this point, we used a dataset named 8D5K, described in [37]. This dataset is composed of five Gaussian clusters of 200 points in dimension 8. Three 2D views of the data were created by selecting three pairs of features for which a clear cluster structure appeared. The fuzzy c-means algorithm (FCM) was applied in each view after selecting by hand the number of desired clusters to obtain three hard partitions computed from the fuzzy partitions. These partitions are represented in Figure 5. The left column shows the partitions in the 2D views, and the right one shows the same partitions projected onto the first two principal components of the data. An ensemble of three mass functions of type I was constructed by considering that the true unknown partition is at least as fine as the individual ones: each clusterer, discounted according the stability of the partition, was represented by a mass function with two focal elements. A “consensus” clustering was obtained by applying the conjunctive rule of combination, computing the matrix $\ln(Pl)$ and the associated tree using Ward linkage, and cutting the tree to obtain five clusters. The consensus clustering and the dendrogram are presented in Figure 6. It may be seen that a clear structure in five clusters is highlighted by the tree and that a very good clustering is obtained. The AR index between the true partition and the result of the ensemble is equal to 0.95.

5.3. Distributed clustering: second experiment

In this section, we illustrate the interest of the discounting process applied to the clusterers according to their stability. We used the same data set as in the previous section and we progressively added to the previous ensemble "noisy" clusterers (from one to nine additional clusterers). Each noisy clusterer was constructed by running the fuzzy c-means algorithm with new pairs of features and by randomly perturbing the labels of the points according to a given noise level (namely 0%, 25%, 50%, 75% and 100%). The results of the ensemble are judged using the Adjusted Rand index between the true partition and the partition into five clusters found by the ensemble. This experiment was repeated 50 times to report errors bars. Our method is compared to the EAC approach.

We can see in Figure 7 that the discount process allows our method to be remarkably stable. On the contrary, the performance of the EAC approach is highly variable and decreases when the number of noisy clusterers added in the ensemble grows.

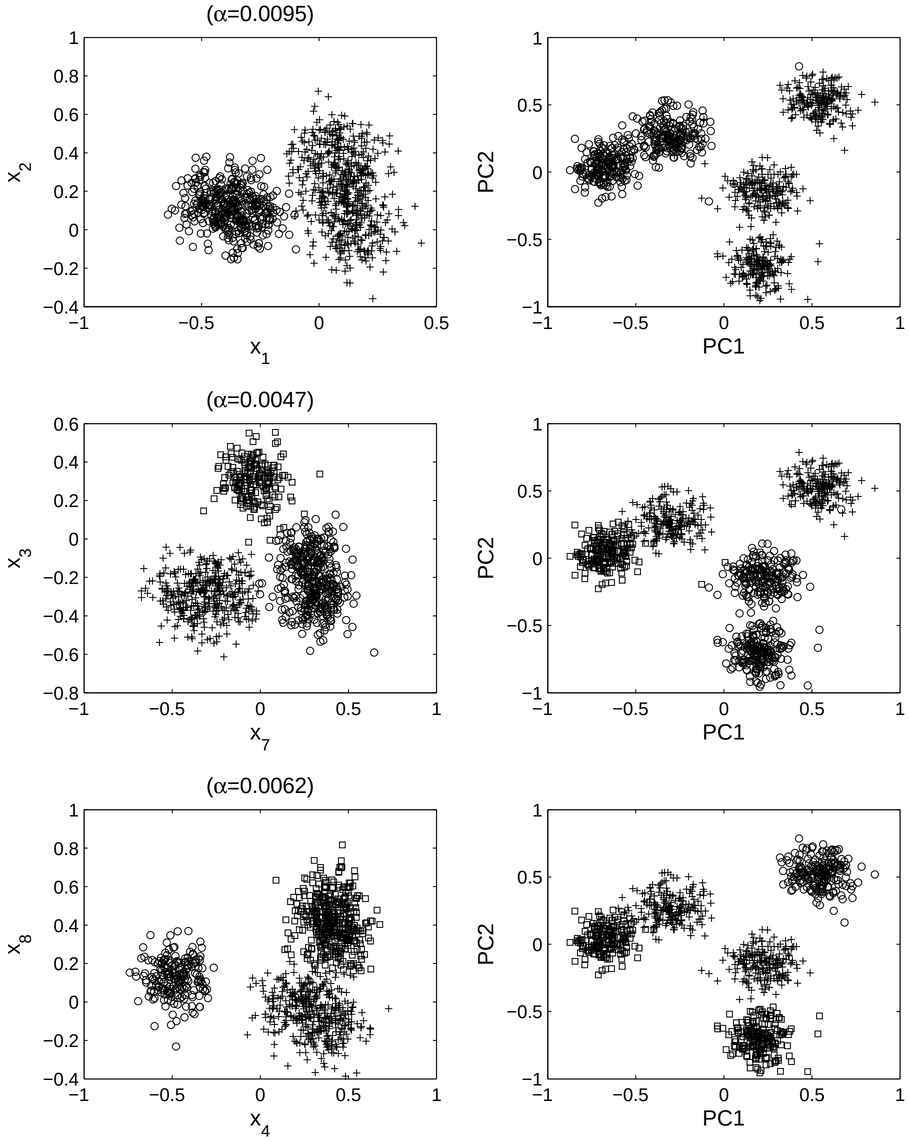

Figure 5: 8D5K data set [37]. The ensemble is composed of three individual clustering solutions obtained from three 2D views of the data. The left column shows the partition obtained in each two-dimensional features space and the right one shows the corresponding partition in the plane spanned by the two first principal components.

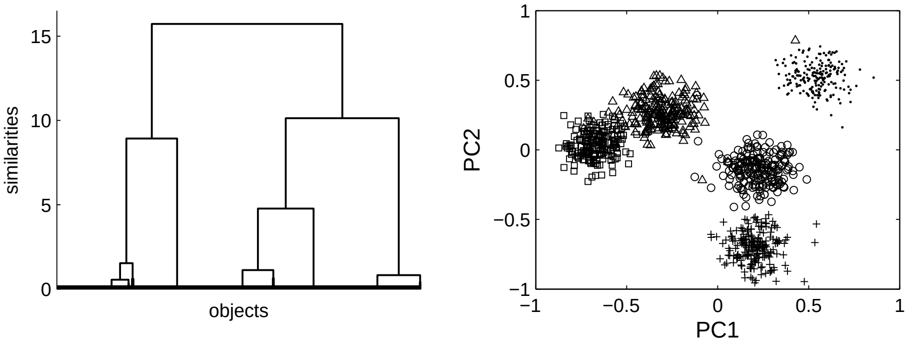

Figure 6: 8D5K data set. Ward's linkage (left) computed from $\ln(Pl)$ and derived consensus partition (right).

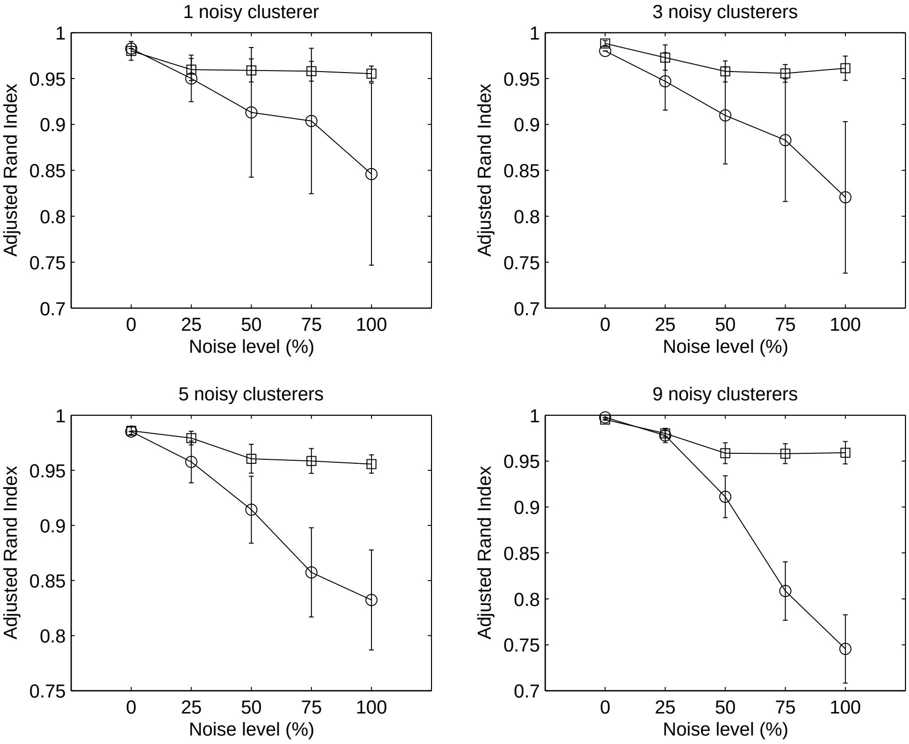

Figure 7: 8D5K data set. Influence of noisy clusterers on the performances of the ensemble (squares: proposed method; circles: EAC approach).

5.4. Discovering non spherical clusters

This section is intended to show the ability of the proposed approach to detect clusters with complex shapes. Two experiments are presented. The first data set is the half-ring data set which is inspired from [14]. It consists of two clusters of 100 points each in a two-dimensional space. To build the ensemble, we used the fuzzy c-means algorithm with a varying number of clusters (from 6 to 11). The hard partitions computed from the soft partitions are represented in Figure 8.

As in the previous example, each partition was discounted according to the stability index and six mass functions with two focal elements each were combined using the conjunctive rule of combination. The mass functions were chosen of type II since the true partition is supposed to be coarser than each individual one. A tree was computed from matrix Bel using Ward's linkage. This tree, represented in the left part of Figure 9, indicates a clear separation in two clusters. Cutting the tree to obtain two clusters gives the partition represented in the right part of Figure 9. We can see that the natural structure of the data is perfectly recovered.

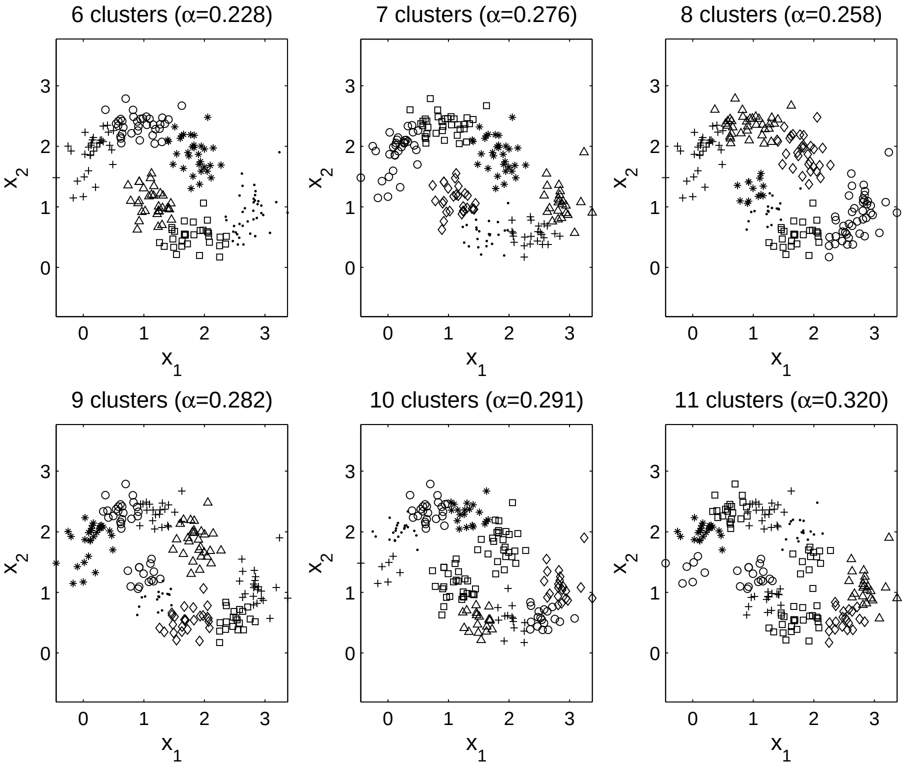

Figure 8: Half-rings data set. Individual partitions.

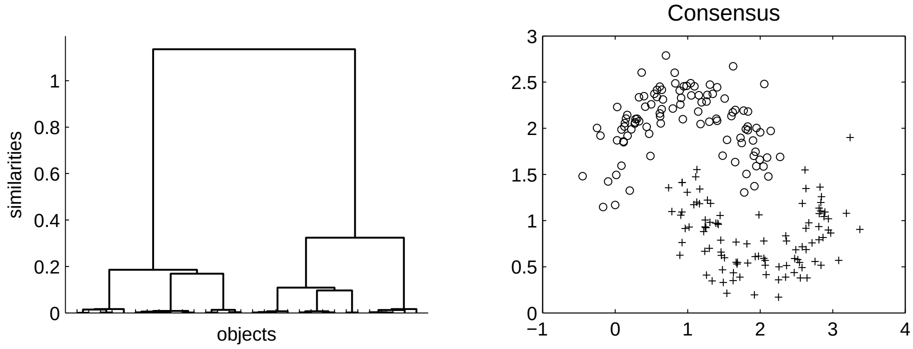

Figure 9: Half-rings data set. Ward's linkage computed from *Bel* and derived consensus.

The second experiment was conducted with the Iris data set, which is a real data set composed of 3 classes of 50 points each in a four dimensional space. This data set is represented in Figure 10. An ensemble of 4 clusterers was constructed using the same approach as before: the fuzzy *c*-means algorithm was run with a varying number of clusters (8 to 11). The corresponding hard partitions were discounted according to the stability criterion and four mass functions of type II were built. These mass function were combined using the conjunctive rule of combination and matrix *Bel* was computed. The corresponding tree obtained using Ward's linkage, which shows a cut in 2 or 3 classes, is represented in the top left of Figure 12. The partition computed from this tree in three classes is shown in the top right of Figure 12. The adjusted Rand index is equal to 0.922.

As a matter of comparison, the tree computed using Ward's linkage from the co-association matrix of the EAC approach and the related partition into three classes are shown in the bottom of Figure 12. It may be seen from the figure that the EAC approach does not give a clear indication about the number of clusters to be chosen and that the partition does not reflect the natural structure of the data (note that the EAC approach and our approach are equivalent only in case of type I assignments). The co-association matrix and the *Bel* matrix are displayed in Figure 12. This representation confirms that the structure of the data is better described by matrix *Bel*.

We also give in Table 1 the averaged Adjusted Rand index and its standard deviation obtained over 100 repetitions of four methods: our method, the EAC approach, and a direct application of a hierarchical clustering (Ward's method) and FCM on the original features. Note that the variability of the results for the first two methods comes, on the one hand, from the variability of the results of FCM (which occurs when the number of clusters is high) and, one the other hand, from the resampling process in the computation of the validity indices.

Belief ens.	EAC approach	Hierarc.	FCM (3 clusters)
0.8821 ± 0.11	0.6139 ± 0.07	0.7312 ± 0	0.7294 ± 0

Table 1: Iris data set. Adjusted Rand Index over 100 repetitions of four approaches.

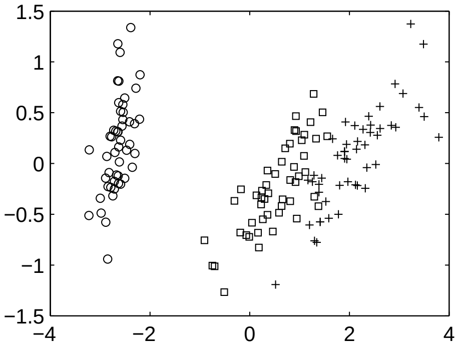

Figure 10: Iris data set. True partition in the plane spanned by the first two principal components.

6. Conclusion

We have proposed in this paper a new approach for aggregating multiple clusterings. This approach is based on the use of belief functions defined on a lattice of sets of partitions. Belief functions theory has been already successfully applied to unsupervised learning problems [23, 7, 25, 26]. In those methods, belief functions are defined on the set of possible clusters, the focal elements being subsets of this frame of discernment. The idea suggested here is radically different. Each clustering algorithm is considered as a source providing an opinion, potentially unreliable, about the unknown partition of the objects. The information of the different sources are converted into masses of evidence allocated to sets of partitions. These masses are then combined and synthesized using some generalizations of classical tools of the belief functions theory. A popular clustering ensemble approach, namely the EAC approach and its weighted version, are recovered as a special case of the method.

Several ways of defining the masses have been suggested in this paper. In particular, type I and type II masses appear as natural expressions of partial knowledge about an unknown partition. We have chosen to relate the masses to a stability index of the partition. This stability index has the advantage of being independent of the feature space and on the clusterer type. Experimental results have shown the ability of the method to recover correct partitions from partial information provided by simple clustering algorithms. Moreover, the robustness of the method against noisy clusterers has been demonstrated in one of the experiments.

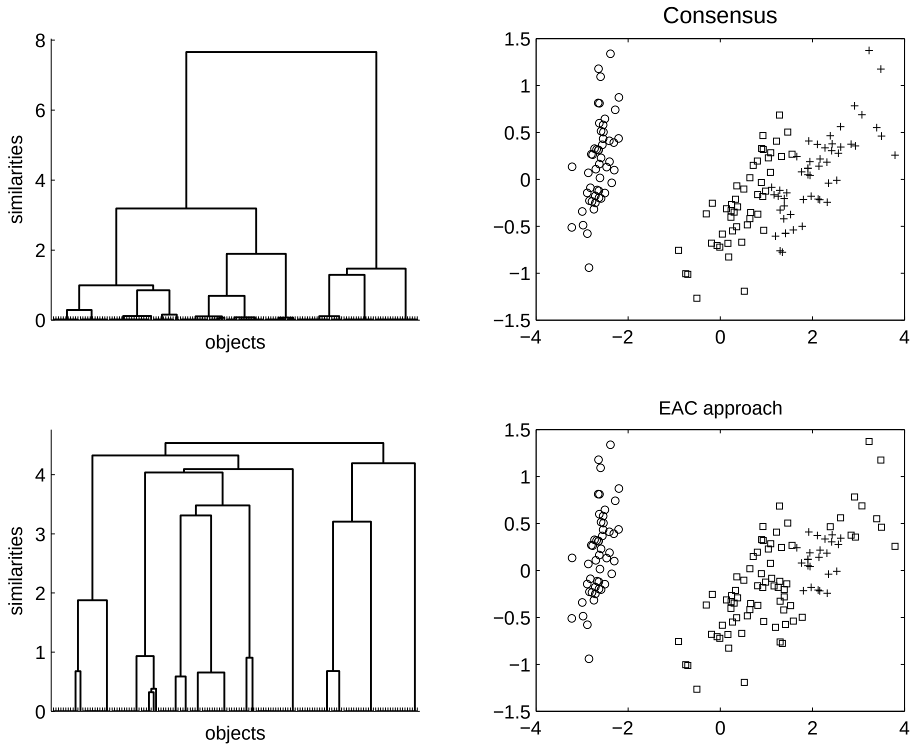

Figure 11: Iris data set. Top: Ward's linkage computed from *Bel* and derived consensus. Bottom: Ward's linkage using the EAC approach and derived consensus.

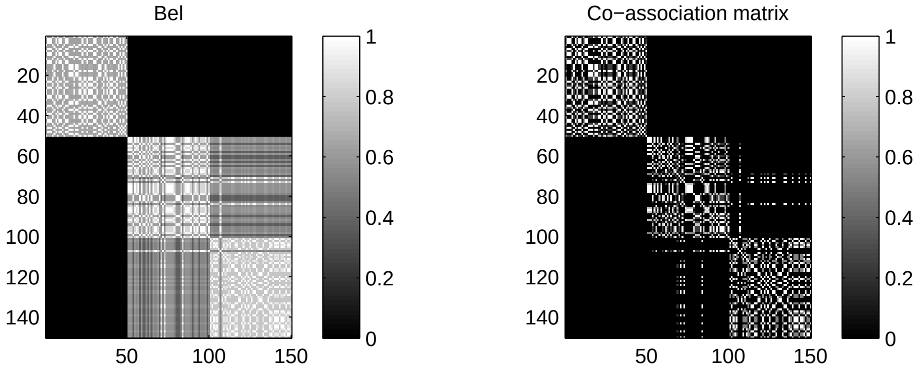

Figure 12: Iris data set. Left: Bel matrix; Right: Co-association matrix computed by the EAC approach.

Appendix

Algorithm 1 CombineEnsemble

```
**Input:** $\mathcal{C} = \{m^1, m^2, ..., m^r\}$
**Output:** $m^*$
$m \leftarrow m^1$
**for** $l = 2$ to $r$ **do**
    $m' \leftarrow m^l$
    $m \leftarrow$ CombineTwoClusterers$(m, m')$;
**end for**
$m^* \leftarrow m$
```

Algorithm 2 CombineTwoClusterers

```
**Input:** Two bbas $m^1 = \{[\underline{p}_{1s}, \overline{p}_{1s}], m_{1s})\}, m^2 = \{[\underline{p}_{2s}, \overline{p}_{2s}], m_{2s})\}$
**Output:** $m^1 \circledcap m^2 = \{[\underline{p}_k, \overline{p}_k], m_k)\}$
$k \leftarrow 1$
**for** $s = 1$ to $n_1$ **do**
  **for** $s' = 1$ to $n_2$ **do**

    % $\underline{p}_k \leftarrow \underline{p}_{1s} \vee \underline{p}_{2s'}$
    $\underline{R}_{p_k} \leftarrow \max(\underline{R}_{1s}, \underline{R}_{2s'})$
    $\underline{R}_{p_k} \leftarrow \text{TransClos}(\underline{R}_{p_k})$

    % $\overline{p}_k \leftarrow \overline{p}_{1s} \wedge \overline{p}_{2s'}$
    $\overline{R}_{p_k} \leftarrow \min(\overline{R}_{1s}, \overline{R}_{2s'})$

    **if** $\overline{R}_p \geq \underline{R}_p$ **then**
            $A_k \leftarrow [\underline{p}_k; \overline{p}_k]$
    **else**
            $A_k \leftarrow \emptyset_{\mathbb{P}_E}$
    **end if**

    $m_k \leftarrow m_{1s} m_{2s'}$
    $k \leftarrow k + 1$

  **end for**
**end for**
```

Algorithm 3 TransClos

```
**Input:** A binary relation $R$
**Output:** T = Transitive closure of $R$
$Continue = True$
$S = R$
**while** $Continue = True$ **do**
  $R \leftarrow R * R$
  $R \leftarrow R > 0$
  **if** $R = S$ **then**
    $Continue = False$
  **end if**
  $S = R$
**end while**
$T = R$
```

Algorithm 4 Validity Index

```
Input: A data set $\mathcal{X}$, a sampling ratio $0.7 \leq S \leq 0.9$, a number $n_{boot}$ of bootstrap samples
Output: C = Cluster Stability Index
for $i = 1$ to $n_{boot}$ do
    Draw two bootstrap samples $\mathcal{X}_1$ and $\mathcal{X}_2$ from $\mathcal{X}$ with a sampling ratio $S$.
    $p_1$ : partition obtained from $\mathcal{X}_1$
    $p_2$ : partition obtained from $\mathcal{X}_2$
    $I = \mathcal{X}_1 \cap \mathcal{X}_1$
    $S(i) \leftarrow \max(0, AR(p_1(I), p_2(I)))$
end for
$C \leftarrow \frac{1}{n_{boot}} \sum_{i=1}^{n_{boot}} S(i)$
```

References

[1] R. Avogadri, and G. Valentini. Fuzzy ensemble clustering for DNA microarray data analysis. *Artificial Intelligence in Medicine*, 45(2-3), 2009.

[2] J.-P. Barthélémy. Monotone functions on finite lattices: an ordinal approach to capacities, belief and necessity functions. In J. Fodor, B. de Baets, and P. Perny, editors, *Preferences and decisions under incomplete knowledge*, Physica-Verlag, 195-208, 2000.

[3] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In Aetman, R.B. (Ed.), In Proc. of the *Pacific Symposium on Biocomputing* , New Jersey World Scientific Publishing Co, 6-17, 2002.

[4] J.C. Bezdek, J. Keller, R. Krisnapuram, and N.R. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Series: The Handbooks of Fuzzy Sets, Kluwer Academic Publishers, Nonwell, MA, USA, Vol. 4, 1999.

[5] D.L. Davies, and D.W. Bouldin. A cluster separation measure. *IEEE Trans. Pattern Anal. Machine Intell.*, 1(4), 224-227, 1979.

[6] W. Day. Foreword: Comparison and Consensus of Classifications. *Journal of Classification*, 3, 183-185, 1986.

[7] T. Denœux, and M.-H. Masson. EVCLUS: EVidential CLUStering of proximity data. *IEEE Transactions on Systems, Man and Cybernetics Part B*, 34(1), 95-109, 2004.

[8] E. Dimitriadou, A. Weingessel, and K. Hornik. Voting-Merging: an ensemble method for clustering. In Proc. of the *International Conference on Artificial Neural Networks*, ICANN'01,, Vienna, Austria, 2001.

[9] F.J. Duarte, and A.L. Fred, and A. Lourenço, and M.F. Rodrigues MF. Weighted evidence accumulation clustering. In: Simoff SJ, Williams GJ, Galloway J, Kolyshkina I (eds), In Proc. of the *4th Australasian Conf Knowl Discovery Data Mining*, Sydney, NSW, Australia, University of Technology, Sydney, 205-220, 2005.

[10] S. Dudoit, and J. Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9), 1090-1099, 2003.

[11] J.C. Dunn. Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics, 4, 95-104, 1974.

[12] X.Z. Fern, and C.E. Broadley. Random projection for high dimensional data clustering: A cluster ensemble approach, Fawcett, T., Mishra, N. (Eds.), In Proc. of the *20th international conference on Machine Learning (ICML 2003)*, AAAI Press, Washington, D.C., USA, 186-193, 2003.

[13] X.Z. Fern, and C.E. Broadley. Solving cluster ensemble problems by bipartite graph partitioning. Brodley CE (ed), In Proc. of the *21th international conference on Machine Learning (ICML 2004)* Banff, AL, Canada. ACM, New York, 281-288, 2004.

[14] A. Fred, and A.K. Jain. Data clustering using evidence accumulation. In Proc. of the *16th International Conference on Pattern Recognition*, Quebec, Canada, 276-280, 2002.

[15] A. Fred, and A.K. Jain. Combining multiple clustering using evidence accumulation. *IEEE Trans Pattern Analysis Mach Intell*, 27, 835-850, 2005.

[16] A. Fred, and A. Lourenço. Cluster Ensemble Methods: from Single Clusterings to Combined Solutions. *Studies in Computational Intelligence (SCI)*, 126, 3-30, 2008.

[17] M. Grabisch. Belief functions on lattices. *International Journal of Intelligent Systems*, 24(1), 76-95, 2009.

[18] D. Greene, A. Tsymbal, N. Bolshakova, and P. Cunningham. Ensemble clustering in medical diagnostics. In Proc. of the *17th IEEE Symposium on Computer-Based Medical Systems*, Bethesda, MD, USA, 576-581, 2004.

[19] S.T. Hadjitodorov, L. Kuncheva, and L. Todorova. Moderate diversity for better cluster ensemble. *Information Fusion*, 7(3), 264-275, 2006.

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

[21] A. Hubert. Comparing partitions. *Journal of Classification*, 2, 193-198, 1985.

[22] A.K. Jain, and R.C. Dubes. *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[23] S. Le Hégarat-Mascle, I. Bloch, and D. Vidal-Madjar. Application of Dempster-Shafer Evidence Theory to Unsupervised Classification in Multisource Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 35(4), 1018-1031, 1997.

[24] E. Levine, and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13, 2573-2593, 2001.

[25] M.-H. Masson, and T. Denœux. Clustering interval-valued data using belief functions. *Pattern Recognition Letters*, 25(2), 163-171, 2004.

[26] M.-H. Masson, and T. Denœux. ECM: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41, 1384-1397, 2008.

[27] M.-H. Masson, and T. Denœux. Belief Functions and Cluster Ensembles. In C. Sossai and G. Chemello (Eds.), ECSQARU 2009, LNAI 5590, Springer-Verlag, 323-334, 2009.

[28] S. Monti, and P. Tamayo, and J.P. Mesirov, and T.R. Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data, *Machine Learning*, 52, 91-118, 2003.

[29] B. Montjardet. The presence of lattice theory in discrete problems of mathematical social sciences. Why. *Mathematical Social Sciences*, 46,103-144, 2003.

[30] D.A. Neumann, and V.T. Norton. On lattice consensus methods. *Journal of Classification*, 3, 225-256, 1986.

[31] W. Rand. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, 66, 846-850, 1971.

[32] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65, 1987.

[33] V. Singh, and L. Mukherjee, and J. Peng, and J. Xu. Ensemble Clustering using Semidefinite Programming. *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 3-6, 2007.

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

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

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

[37] A. Strehl, and J. Ghosh. Cluster ensemble - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3, 563-618, 2002.

[38] A. Topchy, A.K. Jain, and W. Punch. A mixture model for clustering ensembles. In Proc. of the *4th SIAM Conference on Data Mining (SDM'04)*, Lake Buena Vista, USA, 279-390, 2004.

[39] A. Topchy, A.K. Jain, and W. Punch. Clustering Ensembles: Models of Consensus and Weak Partitions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12), 1866-1881, 2005.

[40] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrdl. Constrained k-means clustering with background knowledge. In Proceedings ICML 2001, Williamstown, USA, 577-584, 2001.