Combining clusterings in the belief function framework ☆

Feng Li[a], Shoumei Li[a], Thierry Denœux[b,c,*]

[a] *Beijing University of Technology, College of Applied Sciences, Beijing, China.*
[b] *Université de technologie de Compiègne, CNRS, Heudiasyc (UMR 7253), France.*
[c] *Institut universitaire de France, Paris, France.*

Abstract

In this paper, we propose a clustering ensemble method based on Dempster-Shafer Theory. In the first step, base partitions are generated by evidential clustering algorithms such as the evidential c-means or EVCLUS. Base credal partitions are then converted to their relational representations, which are combined by averaging. The combined relational representation is then made transitive using the theory of intuitionistic fuzzy relations. Finally, the consensus solution is obtained by minimizing an error function. Experiments with simulated and real datasets show the good performances of this method.

Keywords: Evidence theory, Belief functions, Clustering ensemble, Intuitionistic fuzzy relation

1. Introduction

Clustering is an important task in Machine Learning and Pattern Recognition. It is a statistical method to divide objects into groups, in such a way that objects are similar within each group, and dissimilar across different groups. Clustering methods have proved useful in many real-world application domains, such as data mining, image segmentation, etc.

According to the form of clustering output, we can distinguish between hard and soft partitional clustering; the latter includes *fuzzy* and *evidential clustering*. In particular, evidential clustering, based on Dempster-Shafer (DS) theory (also called the theory of belief functions) [31, 8], has recently attracted the attention of many researchers. Evidential clustering computes a *credal partition*, which describes cluster membership uncertainty using DS mass functions.

In recent years, several evidential clustering algorithms have been developed. Denœux and Masson (2004) [12] first introduced an evidential relational clustering method called EVCLUS. This method finds a credal partition such that the degree of conflict between the mass functions associated with any two objects match their dissimilarity. Antoine et al. [3] proposed a constrained version of EVCLUS, called CEVCLUS, which utilizes prior information provided as pairwise constraints. Denœux et al. [14] introduced a faster version of EVCLUS, called k-EVCLUS, where a new cost function is defined and optimized by an iterative row-wise quadratic programming (IRQP)

☆This research was supported by grant No.11571024 from NSFC, and by the Overseas Talent program from the Beijing Government.

*Corresponding author
Email address: tdenoeux@utc.fr (Thierry Denœux)

algorithm. Li et al. [23] further expanded the k-EVCLUS method by taking prior knowledge into account.

In [26], Masson and Denœux introduced the evidential c-means algorithm (ECM), which is an extension of the classic and fuzzy c-means in the framework of DS theory. The ECM alternatively searches for the best credal partition and the best prototypes. Masson et al. [27] proposed a variant of ECM for dissimilarity data, called RECM. Antoine et al. [2] introduced a constrained version of ECM (called CECM) by considering prior knowledge. Liu et al. [25] proposed another variant of the ECM algorithm, called CCM, by introducing the notion of meta-cluster. Zhou et al. [43] extended the median c-means and median fuzzy c-means to the Median Evidential c-means (MECM). Denœux et al. (2015) [10] introduced a new evidential clustering algorithm (Ek-NNclus) based on the evidential k nearest neighbor rule.

Different clustering algorithms may obtain different clustering results for one dataset, and even a single algorithm with different initializations may yield different solutions. It is generally agreed that there is no best single clustering algorithm [1]. To solve this problem and further improve the robustness, consistency and stability of the solution, clustering ensemble methods have emerged as an approach for combining multiple clustering results into an improved solution. Among those, the Evidence Accumulation Clustering (EAC) method [16, 17] has attracted a lot of attention. It constructs a co-association matrix from base partitions, which can be processed by a hierarchical clustering algorithm to obtain the final solution.

In this paper, we introduce an evidential clustering ensemble method that can be seen as an extension of the EAC method in the DS framework. In our method, base evidential partitions are generated by evidential clustering algorithms. As noted in [14], evidential partitions allow for ambiguity, uncertainty or doubt in the assignment of objects to clusters, and constitute a rich and informative description of the clustering structure of a dataset. The base credal partitions are transformed into their relational representations [11], which are combined using different rules of DS theory. The fused relational representation corresponds to the co-association matrix in the EAC method. It can be seen as defining an intuitionistic fuzzy relation [22], which is made transitive to obtain an intuitionistic fuzzy equivalence relation. The final evidential partition is obtained by minimizing an objective function based on the credal Rand index [11]. Experiments with real and simulated data show that our approach can reveal the underlying clustering structure of complex-shape datasets and achieve better results than EAC and single clustering algorithms.

The rest of this paper is organized as follows. Basic notions are first recalled in Section 2. Our method is then described in Section 3, and experimental results are reported in Section 4. Finally, Section 5 concludes the paper.

2. Background

In this section, we briefly introduce some basic notions used in this paper. The main definitions of DS theory are first recalled in Section 2.1. The notions of credal partition and relational representation are then reviewed in Section 2.2. Section 2.3 introduces intuitionistic fuzzy relations and the transitive closure theorem. We then describe the EAC method in Section 2.4, and we review some related work in Section 2.5.

2.1. DS theory

Let $\Omega = \{\omega_1, \ldots, \omega_c\}$ be a finite set. A *mass function* on Ω is a mapping from the power set 2^Ω to $[0, 1]$, satisfying the condition

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

Each subset A of Ω such that $m(A) > 0$ is called a *focal set*. In DS theory, a mass function is viewed as a piece of evidence about some question of interest, for which the true answer, denoted by ω, is supposed to be an element of Ω. For any nonempty focal set A, $m(A)$ is a measure of the belief that is committed exactly to A [31]. The mass $m(\emptyset)$ assigned to the empty set is a measure of the belief that the true answer might not belong to Ω. The mass $m(\Omega)$ is a measure of ignorance. A mass function is said to be *logical* if it has only one focal set. It is said to be *normalized* if the empty set is not a focal set, and *unnormalized* otherwise. An unnormalized mass function m can be converted into a normalized one m^* by *Dempster's normalization operation* defined by $m^*(\emptyset) = 0$ and

$$m^*(A) = \frac{m(A)}{1 - m(\emptyset)}. \tag{2}$$

Given a mass function m, the corresponding *belief* and *plausibility* functions are defined, respectively, as

$$bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B)$$

and

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

for all $A \subseteq \Omega$. Clearly, functions *bel* and *pl* are linked by the relation $pl(A) = 1 - bel(\overline{A})$, where $\overline{A}$ denotes the complement of A. The quantity $bel(A)$ represents the degree of total support in A, while $pl(A)$ can be interpreted as the degree to which the evidence is consistent with A.

Different combination rules have been proposed in the literature. For example, the *conjunctive rule* [32] and the dual *disjunctive rule* [33] are defined, respectively, as

$$(m_1 \textcircled{\cap} m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C), \tag{3a}$$

$$(m_1 \textcircled{\cup} m_2)(A) = \sum_{B \cup C = A} m_1(B) m_2(C), \tag{3b}$$

for any two mass functions m_1 and m_2 on the same frame Ω and all $A \subseteq \Omega$. As shown by Smets [33], the conjunctive rule assumes that all mass functions to be combined are derived from reliable sources of information, whereas the disjunctive rule only assumes that at least one source of information is reliable, but we do not know which one. Dempster's rule, denoted by $\oplus$, is defined by the conjunctive rule (3a) followed by normalization (2), i.e., $(m_1 \oplus m_2)(\emptyset) = 0$ and

$$(m_1 \oplus m_2)(A) = \frac{(m_1 \textcircled{\cap} m_2)(A)}{1 - (m_1 \textcircled{\cap} m_2)(\emptyset)} \tag{4}$$

for all $A \neq \emptyset$.

The *Dubois-Prade (DP) rule* [15] assumes that when two sources are not in conflict, they are

both reliable, and at least one is correct when they are conflicting. Specifically, for two mass functions m_1 and m_2, their combination by the DP rule, denoted by $m_1 \boxplus m_2$, is

$$(m_1 \boxplus m_2)(A) = \sum_{\substack{B\cap C=A \\ B\cap C\neq\emptyset}} m_1(B)m_2(C) + \sum_{\substack{B\cup C=A \\ B\cap C=\emptyset}} m_1(B)m_2(C), \quad \forall A \neq \emptyset. \tag{5}$$

It can be seen as a reasonable trade-off between the conjunctive and disjunctive rules.

Let us now assume that mass function m represents our current state of knowledge about ω, and we need to choose one or several elements of Ω as our estimate about the true answer. Decision rules in the DS framework are reviewed in [7]. Here, we only mention two rules that will be used in the sequel. The *maximum plausibility rule* selects the element ω^* with the highest plausibility,

$$\omega^* = \arg\max_{\omega\in\Omega} pl(\{\omega\}). \tag{6}$$

This rule yields one single result. In contrast, the *interval dominance rule* [7] is based on the following dominance relation: ω dominates ω', iff $bel(\{\omega\}) > pl(\{\omega'\})$. Then the set of its maximal (non-dominated) elements can be obtained as

$$\Omega^* = \{\omega \in \Omega \mid pl(\{\omega\}) \geq bel(\{\omega'\}), \forall \omega' \in \Omega\}. \tag{7}$$

Instead of reaching a single decision, this rule selects a set of potential results.

2.2. Credal partitions

Let $\mathcal{O}$ be a set of n objects, and $\Omega = \{\omega_1, ..., \omega_c\}$ the set of clusters. Each object is assumed to belong to at most one cluster. Uncertain knowledge about the cluster-membership of object o_i is represented by a mass function m_i on Ω. A *credal partition* [12] is defined as an n-tuple $M = (m_1, .., m_n)$. The notion of credal partition is more general than those of hard or fuzzy partitions, and a credal partition can be summarized into a partition of any other type [9]. Credal partitions also encompass rough partitions as a special case [30, 14]: a rough partition corresponds to a credal partition in which all mass functions m_i are logical, namely $m_i(A_i) = 1$ for some $A_i \subseteq \Omega$. The *lower and upper approximations* of cluster ω_k can then be defined as follows,

$$\omega_k^L = \{o_i \in \mathcal{O} \mid A_i = \{\omega_k\}\} \tag{8a}$$

and

$$\omega_k^U = \{o_i \in \mathcal{O} \mid \omega_k \in A_i\}. \tag{8b}$$

For a non-logical mass function m_i, the set A_i can be selected by the interval dominance rule (7), after which a rough partition can be obtained.

Suppose m_i and m_j are two mass functions related to objects i and j. We consider the frame $\Theta_{ij} = \{s_{ij}, \neg s_{ij}\}$, where s_{ij} means that "Objects i and j belong to the same cluster", and $\neg s_{ij}$ means that "Objects i and j belong to different clusters". A mass function m_{ij} on Θ_{ij} representing our beliefs about the joint cluster-membership of objects i and j can be computed from m_i and

m_j as follows [11]:

$$m_{ij}(\emptyset) = m_i(\emptyset) + m_j(\emptyset) - m_i(\emptyset)m_j(\emptyset), \tag{9a}$$

$$m_{ij}(\{s_{ij}\}) = \sum_{k=1}^{c} m_i(\{\omega_k\})m_j(\{\omega_k\}), \tag{9b}$$

$$m_{ij}(\{\neg s_{ij}\}) = \sum_{A\cap B=\emptyset} m_i(A)m_j(B) - m_{ij}(\emptyset), \tag{9c}$$

$$m_{ij}(\Theta_{ij}) = \sum_{A\cap B\neq\emptyset} m_i(A)m_j(B) - m_{ij}(\{s_{ij}\}), \tag{9d}$$

The tuple $\mathcal{R} = (m_{ij})_{1\leq i\leq j\leq n}$ is called the *relational representation* of M. We note that $m_{ij}(\emptyset) = 0$ whenever $m_i(\emptyset) = 0$ and $m_j(\emptyset) = 0$. Given two credal partitions M and M' and their relational representations $\mathcal{R}$ and $\mathcal{R}'$, the *credal Rand index* [11] is defined as

$$\rho_S(\mathcal{R}, \mathcal{R}') = 1 - \frac{\sum_{i<j} \delta(m_{ij}, m'_{ij})}{n(n-1)/2}, \tag{10}$$

where δ is the Jousselme's distance [20]. In the special case where mass functions are normal, i.e., $m_{ij}(\emptyset) = m'_{ij}(\emptyset) = 0$, we can write mass function m_{ij} as a vector

$$\mathbf{m}_{ij} = (m_{ij}(\{s_{ij}\}), m_{ij}(\{\neg s_{ij}\}), m_{ij}(\Theta_{ij}))^T.$$

Jousselme's distance between m_{ij} and m'_{ij} is then defined as

$$\delta(m_{ij}, m'_{ij}) = \Big(\frac{1}{2}(\mathbf{m}_{ij} - \mathbf{m}'_{ij})^T\mathbf{J}(\mathbf{m}_{ij} - \mathbf{m}'_{ij})\Big)^{1/2},$$

where $\mathbf{J}$ is the Jaccard matrix

$$\mathbf{J} = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{pmatrix}. \tag{11}$$

The range of ρ_S is $[0, 1]$, and it boils down to the Rand index when both M and M' are hard partitions. The credal Rand index measures the similarity between any two soft partitions.

2.3. Intuitionistic Fuzzy Equivalence Relation

Fuzzy relations. In classical set theory, an equivalence relation provides a partition of the underlying set into disjoint equivalence classes. In fuzzy set theory, a *fuzzy relation* R on a finite set X is defined as a fuzzy subset of the Cartesian product X^2, i.e., a mapping from X^2 to $[0, 1]$ [29]. Each membership value $R(x, y)$ represents the degree to which x stands in relation R with y. For a fuzzy relation R, we define the following properties:

Reflexivity: For all $x \in X$, $R(x, x) = 1$;

Irreflexivity: For all $x \in X$, $R(x, x) = 0$;

Symmetry: For all $(x, y) \in X^2$, $R(x, y) = R(y, x)$;

Transitivity: For all $(x, y, z) \in X^3$, $\min(R(x, y), R(y, z)) \leq R(x, z)$.

Furthermore, R is *dual transitive* iff $1 - R$ is transitive. If a fuzzy relation is reflexive, symmetric and transitive, it is called a *fuzzy equivalence relation.*

Sometimes, another triangular norm (t-norm) than the minimum is used in the definition of transitivity. For a t-norm T, R is said to be *T-transitive* if for all $(x, y, z) \in X^3$,

$$T\left(R(x, y), R(y, z)\right) \leq R(x, z).$$

The *max-T composition* for two fuzzy relations R and Q on X is the fuzzy relation $R \circ Q$ defined by

$$(R \circ Q)(x, y) = \max_{z \in X} T(R(x, z), Q(z, y)). \tag{12}$$

Denoting $R \circ R$ as R^2, the T-transitivity property can be expressed as $R^2 \subseteq R$.

The max-T transitive closure $\overline{R}$ of a fuzzy relation R is the smallest max-T transitive fuzzy relation containing R. It can be computed as

$$\overline{R} = \bigcup_{i=1}^{\infty} R^i, \tag{13}$$

where $\cup$ is the fuzzy set union based on the maximum t-conorm. In particular, if R is a reflexive and symmetric fuzzy relation on a finite set X of cardinality n, then $\overline{R} = R^{n-1}$ [4].

Intuitionistic Fuzzy relations. The notion of *Intuitionistic Fuzzy Relation* (IFR) is a further generalization of relations based on the theory of intuitionistic fuzzy sets [22]. An intuitionistic fuzzy subset (IFS) A of X is a pair of mappings $\mu_A : X \to [0, 1]$ and $\nu_A : X \to [0, 1]$ such that $\mu_A(x) + \nu_A(x) \leq 1$ for each $x \in X$. The values $\mu_A(x)$ and $\nu_A(x)$ represent, respectively, the membership degree and non-membership degree of element x in the set A. The pair $(\mu_A(x), \nu_A(x))$ is called an intuitionistic fuzzy value (IFV).

Let L be the set of all IFVs, i.e., $L = \{\alpha = (\mu_\alpha, \nu_\alpha) \mid \mu_\alpha \in [0, 1], \nu_\alpha \in [0, 1], \mu_\alpha + \nu_\alpha \leq 1\}$. A partial ordering relation on L can be defined as follow:

$$\alpha \leq_L \alpha' \Leftrightarrow \mu_\alpha \leq \mu_{\alpha'} \quad \text{and} \quad \nu_\alpha \geq \nu_{\alpha'},$$

for all $(\alpha, \alpha') \in L^2$. Any pair $(\alpha, \alpha') \in L^2$ has a unique least upper bound $\alpha \vee \alpha'$ and a unique greatest lower bound $\alpha \wedge \alpha'$ given, respectively, by

$$\alpha \vee \alpha' = \big(\max(\mu_\alpha, \mu_{\alpha'}), \min(\nu_\alpha, \nu_{\alpha'})\big)$$

and

$$\alpha \wedge \alpha' = \big(\min(\mu_\alpha, \mu_{\alpha'}), \max(\nu_\alpha, \nu_{\alpha'})\big).$$

Thus $(L, \leq_L)$ is a complete lattice, with top $(1, 0)$ and bottom $(0, 1)$.

An IFR on a non-empty set X is an IFS of X^2, i.e., a mapping $R : X^2 \to L$. For an IFR R we define the following properties:

Reflexivity: For all $x \in X$, $R(x, x) = (1, 0)$;

Symmetry: For all $(x, y) \in X^2$, $R(x, y) = R(y, x)$;

Transitivity: For all $(x, y, z) \in X^3$, $\bigvee_{y \in X}(R(x, y) \wedge R(y, z)) \leq_L R(x, z)$.

A reflexive, symmetric and transitive IFR is called an *Intuitionistic Fuzzy Equivalence Relation* (IFER). An IFR $R = (\mu_R, \nu_R)$ is an IFER if and only if μ_R is reflexive, symmetric and transitive, and ν_R is irreflexive, symmetric and dual transitive [22]. Consequently, the transitivity of an IFR can be obtained by making μ_R and $1 - \nu_R$ transitive. In the experiments reported in Section 4, we consider three t-norms: the minimum, the product, and the Lukasiewicz t-norm defined as $T(a, b) = \max(0, a + b - 1)$.

2.4. Evidence Accumulation Clustering

In this section, we briefly summarize the EAC method [16, 17]. More details can be found in [18, 38, 42]. The EAC method uses the *co-association matrix* to avoid the label correspondence problem. More precisely, assume that a dataset has n objects $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$. Suppose that N base partitions $P^1, \ldots, P^N$ have been obtained in the first step. In the second step, each base partition P^b is mapped to a co-association matrix S^b of size $N \times N$ with general term $s^b_{ij} = I(c^b_i = c^b_j)$, where c^b_i is the cluster index of x_i in P^b and I is the indicator function. The co-association matrix, denoted as $S^* = (s^*_{ij})$, is the average of all S^b; its general term is

$$s^*_{ij} = \frac{1}{N} \sum_{b=1}^{N} s^b_{ij}. \tag{14}$$

Each element s^*_{ij} represents the proportion of base partitions in which objects o_i and o_j are assigned to the same cluster. The co-association matrix can be treated as a new similarity matrix and used as input to single-linkage hierarchical clustering.

In EAC, the co-association matrix is computed by only taking into account whether two objects belong to the same cluster or not. Some researchers [19, 37, 40] have proposed to use additional information to construct a similarity measure that is more expressive about the relationship between objects. For instance, Yang [41] proposed a fuzzy co-association matrix to summarize the ensemble of fuzzy partitions, where the membership of an object to clusters is expressed by a fuzzy membership function.

We can remark that, in the EAC method, the numbers of clusters in the base partitions do not need to be close to the "true" number of clusters. Indeed, in [17], the authors construct the base partitions using the k-means algorithms with a large (and sometimes random) number of clusters. For instance, in one dataset with two clusters, they construct base partitions with up to 80 clusters. The underlying assumption is that objects that truly belong to true same cluster are likely to be assigned to the same cluster in different partitions, which is summarized in the co-association matrix. The number of clusters in the final aggregated partition is determined by analyzing the dendrogram after applying single-link hierarchical clustering to the co-association matrix.

In this paper, we propose to extend the EAC method in the framework of DS theory. In our method, the membership of an object to clusters is represented by a mass function in the credal partition, which contains more information than fuzzy and hard partitions. To better exploit this type of information, we use the relational interpretation recalled in Section 2.2 to measure the "similarity" between objects, which can be seen as a generalization of the co-association matrix. To capture the neighborhood relationship, we make the combined relational representation transitive based on IFR theory recalled in Section 2.3.

2.5. Other related Work

In the ensemble clustering literature, most contributions focus on hard partitions, and some are based on fuzzy partitions. Only a few methods are based on DS theory. DS theory is a sound approach for ensemble clustering methods, because it provides ways to combine different pieces of evidence. Preliminary results have already demonstrated the feasibility of this approach. For example, in [13, 28] the authors propose to define mass functions on the lattice of interval partitions of a set of objects; they obtain a consensus belief function by a suitable combination rule, and use hierarchical clustering to get the final partition. Unlike other direct methods relying on a voting process, Li et al. [24] introduce another direct approach based on Dempster's rule of combination; their method consists of two steps: finding the correspondence labels and using the combination rule to produce the final result. Although these ensemble clustering methods are rooted in Dempster's theory, they still consider hard partitions as input and also output of the procedure. Wang et al. [39] propose an ensemble clustering method for evidential partitions. After solving the label correspondence problem, the final results are obtained by combining the selected base partitions. Due to high computational complexity, they only consider the fixed (true) number of clusters in the base partitions.

In this paper, we combine credal partitions, a very general form of partitions that can be generated by hard, fuzzy, rough or evidential clustering algorithms. After mapping the base partitions to their relational representations, we combine these partitions in a coarser frame, where we only need to consider two focal sets. The combined relational representation is made transitive using the theory of IFRs. Finally, we generate an evidential partition that matches the combined relational representation, providing a much more informative output than can be obtained using hard or fuzzy clustering methods.

3. Evidential Clustering Ensemble Method

In this section, we introduce the proposed method. The generation and combination of base credal partitions are described in Section 3.1. The computation of transitive closures to make the combined relational representation transitive is then addressed in Section 3.2, and the method for computing the final credal partition is presented in Section 3.3.

3.1. Generation of base partitions

In the first step of our method, base partitions can be obtained by hard, fuzzy and rough clustering methods, which all produce special forms of credal partitions. In this paper, we focus on base partitions generated by evidential clustering methods. We assume that we start with N base partitions $M^b = (m_1^b, \ldots, m_n^b)$, $b = 1, ..., N$ generated by evidential clustering algorithms such as ECM or EVCLUS. The number of clusters in base partition b is denoted by k_b. Before converting credal partitions to their relational representations, we compute the average mass assigned to the empty set for each object i as

$$\widehat{m}_i(\emptyset) = \frac{1}{N}\sum_{b=1}^{N} m_i^b(\emptyset), \quad i = 1, ..., n, \tag{15}$$

and we normalize each base credal partition by (2). We denote the b-th normalized credal partition as $M^{b*} = (m_1^{b*}, \ldots, m_n^{b*})$, where m_i^{b*} is the normalized mass function defined by $m_i^{b*}(\emptyset) = 0$ and

$$m_i^{b*}(A) = \frac{m_i^b(A)}{1 - m_i^b(\emptyset)}, \tag{16}$$

for all nonempty subset A of Ω. The reason for this normalization is that the analogy between relation representations and IFRs developed in Section 3.2 requires the mass functions to be normalized. However, we cannot just discard the mass on the empty set, because it is useful to detect outliers. This information will be utilized in the last step of our method (see Section 3.3).

After the relational representations $\mathcal{R}^{b*}$ have been computed, they can be combined using different rules. The combined relational representation is denoted by $\mathcal{R}^* = (m_{ij}^*)_{1 \leq i \leq j \leq n}$. For example, using the average rule we get

$$m_{ij}^* = \frac{1}{N} \sum_{b=1}^{N} m_{ij}^{b*}.$$

We can remark that, when the base partitions are hard, the $n \times n$ matrix with general term $[m_{ij}^*(\{s\})]$ boils down to the co-association matrix (14). As a consequence, EAC is a special case of our method. Other combination rules introduced in Section 2.1 can also be used.

3.2. Transitivity

The combined relational representation $\mathcal{R}^*$ sometimes cannot be successfully exploited because it lacks a notion of transitivity, namely: if we believe that objects i and j belong to the same cluster, and objects j and k also belong to the same cluster, then we should believe that this is also the case for objects i and k. When applying hierarchical clustering in the second step, this property is used implicitly in the EAC method.

A relational representation is not a classical fuzzy relation, because it specifies two numbers for each pair (i, j); the degree of belief that i and j belong to the same cluster, and the degree of belief that they do not belong to the same cluster. We can observe the similarity between this kind of information and IFRs recalled in Section 2.3. Using this formal analogy, we consider the combined relational representation $\mathcal{R}^*$ as an IFR, and we make it transitive using the techniques reviewed in Section 2.3.

For a normalized mass function m_{ij}^* on Θ_{ij}, the degrees of belief in s_{ij} and $\neg s_{ij}$ are, respectively,

$$bel_{ij}^*(\{s_{ij}\}) = m_{ij}^*(\{s_{ij}\})$$

and

$$bel_{ij}^*(\{\neg s_{ij}\}) = m_{ij}^*(\{\neg s_{ij}\}),$$

with $bel_{ij}^*(\{s_{ij}\}) + bel_{ij}^*(\{\neg s_{ij}\}) \leq 1$. These numbers define an IFR $R = (\mu_R, \nu_R)$ on the set $\mathcal{O}$ of objects, with

$$\mu_R(o_i, o_j) = bel_{ij}^*(\{s_{ij}\})$$

and

$$\nu_R(o_i, o_j) = bel_{ij}^*(\{\neg s_{ij}\}).$$

This IFR is reflexive (as $bel_{ii}^*(\{s_{ii}\}) = 1$ and $bel_{ii}^*(\{\neg s_{ii}\}) = 0$ for all i) and symmetric, but it is

usually not transitive. As recalled in Section 2.3, R is transitive iff μ_R is transitive and ν_R is dual transitive, i.e., $\nu_R^d = 1 - \nu_R$ is transitive. To obtain an IFER, we thus need to replace μ_R and ν_R^d by their transitive closures. We observe that

$$\nu_R^d(o_i, o_j) = 1 - bel_{ij}^*(\{\neg s_{ij}\}) = pl_{ij}^*(\{s_{ij}\}).$$

Denoting by $\overline{\mu_R}$ and $\overline{\nu_R^d}$ the transitive closures of μ_R and ν_R^d, let

$$\overline{bel}_{ij}(\{s_{ij}\}) = \overline{\mu_R}(o_i, o_j)$$

and

$$\overline{pl}_{ij}(\{s_{ij}\}) = \overline{\nu_R^d}(o_i, o_j).$$

These belief and plausibility values correspond to new mass functions $\overline{m}_{ij}$ such that $\overline{m}_{ij}(\emptyset) = 0$ and

$$\overline{m}_{ij}(\{s_{ij}\}) = \overline{bel}_{ij}(\{s_{ij}\}), \tag{17a}$$

$$\overline{m}_{ij}(\{\neg s_{ij}\}) = 1 - \overline{pl}_{ij}(\{s_{ij}\}), \tag{17b}$$

$$\overline{m}_{ij}(\Theta_{ij}) = \overline{pl}_{ij}(\{s_{ij}\}) - \overline{bel}_{ij}(\{s_{ij}\}). \tag{17c}$$

The new relational representation $\overline{\mathcal{R}} = (\overline{m}_{ij})_{1 \leq i \leq j \leq n}$ will hereafter be referred to as the *transitivized combined relational representation.*

3.3. Recovering the combined credal partition

In the EAC method, the final result is obtained by applying hierarchical clustering to the association matrix. We could apply the same procedure to matrices $(\overline{bel}_{ij}(\{s_{ij}\}))$ or $(\overline{pl}_{ij}(\{s_{ij}\})$, similarly to what was proposed in [28]. However, our objective is to recover a credal partition, which cannot be obtained by a hierarchical clustering algorithm.

Our approach is to find a normalized credal partition M^* whose relation representation is as close as possible to the transitivized combined relational representation $\overline{\mathcal{R}}$, closeness being measured by the credal Rand index (10). We need to fix the number k_r of clusters as well as the focal sets in the credal partition M^*. For example, we can take the singletons and Ω as focal sets, or we can also consider some pairs of clusters.

We thus seek a normal credal partition M^*, solution of the maximization problem

$$\max_{M^*} \rho_S(\mathcal{R}(M^*), \overline{\mathcal{R}}), \tag{18}$$

where $\mathcal{R}(M^*)$ is the relational representation of M^*. From (10), maximizing (18) is equivalent to minimizing the following stress (error) function:

$$S(M^*) = \sum_{i<j} (\mathbf{m}_{ij}^* - \overline{\mathbf{m}}_{ij})^T \mathbf{J} (\mathbf{m}_{ij}^* - \overline{\mathbf{m}}_{ij}), \tag{19}$$

where

$$\mathbf{m}_{ij}^* = (m_{ij}^*(\{s_{ij}\}), m_{ij}^*(\{\neg s_{ij}\}), m_{ij}^*(\Theta_{ij}))^T,$$
$$\overline{\mathbf{m}}_{ij} = (\overline{m}_{ij}(\{s\}), \overline{m}_{ij}(\{\neg s\}), \overline{m}_{ij}(\Theta_{ij}))^T,$$

and $\mathbf{J}$ is the Jaccard matrix (11). To solve (19), let us write $\mathbf{m}^*_{ij}$ as a function of mass functions m^*_i and m^*_j in matrix form. Assuming that each mass function m^*_i has f focal sets $F_1, \ldots, F_f$, it can be written as an f-vector $\mathbf{m}^*_i = (m^*_i(F_1), \ldots, m^*_i(F_f))^T$, and the credal partition can be written as an $n \times f$ matrix $\mathbf{M}^* = (\mathbf{m}^*_1, \ldots, \mathbf{m}^*_n)^T$. Let $\mathbf{S} = (S_{k\ell})$ and $\mathbf{C} = (C_{kl})$ be the $f \times f$ matrices defined as follows:

$$S_{kl} = \begin{cases} 1 & \text{if } k = l \text{ and } |F_k| = 1 \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

and

$$C_{kl} = \begin{cases} 1 & \text{if } F_k \cap F_l = \emptyset, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

We have:

$$m^*_{ij}(\{s_{ij}\}) = (\mathbf{m}^*_i)^T \mathbf{S}\, \mathbf{m}^*_j, \tag{22a}$$

$$m^*_{ij}(\{\neg s_{ij}\}) = (\mathbf{m}^*_i)^T \mathbf{C}\, \mathbf{m}^*_j, \tag{22b}$$

$$m^*_{ij}(\Theta_{ij}) = 1 - m^*_{ij}(\{s_{ij}\}) - m^*_{ij}(\{\neg s_{ij}\}). \tag{22c}$$

We can observe that $\mathbf{m}^*_{ij}$ is linear in m^*_i and, consequently, $S(M^*)$ is quadratic in m^*_i, the other mass functions being fixed. Consequently, we can minimize $S(M^*)$ using a cyclic coordinate descent algorithm, minimizing $S(M^*)$ with respect to each $\mathbf{m}^*_i$ in turn, while keeping the other $\mathbf{m}^*_j$ constant; this is the iterative row-wise quadratic programming (IRQP) algorithm [35], also used in [14]. Using this approach, we minimize at each step the following cost function:

$$g_i(\mathbf{m}^*_i) = \sum_{\substack{j=1 \\ j \neq i}}^{n} (\mathbf{m}^*_{ij} - \overline{\mathbf{m}}_{ij})^T J (\mathbf{m}^*_{ij} - \overline{\mathbf{m}}_{ij}), \tag{23}$$

which is quadratic in $\mathbf{m}^*_i$. To simplify the expression of this function, let us define the matrix $\mathbf{A}_j$ of size $3 \times 3f$ as

$$\mathbf{A}_j = \mathbf{I}_3 \otimes (\mathbf{m}^*_j)^T. \tag{24}$$

where $\mathbf{I}_3$ is the identity matrix of size 3×3 and $\otimes$ is the Kronecker product, and the matrix $\mathbf{B}$ of size $3f \times f$ as

$$\mathbf{B} = \begin{pmatrix} \mathbf{S} \\ \mathbf{C} \\ 1 - \mathbf{S} - \mathbf{C} \end{pmatrix}. \tag{25}$$

With these notations, (23) can be written as

$$g_i(\mathbf{m}_i) = \sum_{\substack{j=1 \\ j \neq i}}^{n} (\mathbf{A}_j \mathbf{B} \mathbf{m}^*_i - \overline{\mathbf{m}}_{ij})^T \mathbf{J} (\mathbf{A}_j \mathbf{B} \mathbf{m}^*_i - \overline{\mathbf{m}}_{ij}) \tag{26}$$

Developing the right-hand side of (26) and rearranging the terms, we obtain

$$g_i(\mathbf{m}^*_i) = (\mathbf{m}^*_i)^T \mathbf{\Sigma} \mathbf{m}^*_i + \mathbf{u}^T \mathbf{m}^*_i + c_0, \tag{27}$$

with

$$\boldsymbol{\Sigma} = \mathbf{B}^T \left(\sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbf{A}_j^T \mathbf{J} \mathbf{A}_j \right) \mathbf{B} \tag{28a}$$

$$\mathbf{u} = -2 \left(\sum_{\substack{j=1 \\ j \neq i}}^{n} \overline{\mathbf{m}}_{ij}^T \mathbf{J} \mathbf{A}_j \right) \mathbf{B} \tag{28b}$$

$$c_0 = \sum_{\substack{j=1 \\ j \neq i}}^{n} \overline{\mathbf{m}}_{ij}^T \mathbf{J} \overline{\mathbf{m}}_{ij}. \tag{28c}$$

Minimizing $g(\mathbf{m}_i)$ under the constraints $(\mathbf{m}_i^*)^T \mathbf{1} = 1$ (where $\mathbf{1} = (1, \ldots, 1)^T$) and $\mathbf{m}_i^* \geq \mathbf{0}$ is a quadratic programming (QP) problem, which can be solved efficiently with any QP solver. As we iteratively update each row of $\mathbf{M}^*$, the overall stress $S(M^*)$ decreases and eventually reaches a local minimum. As in [14], we compute the following running mean after each cycle of the algorithm as $e_0 = 1$ and

$$e_t = 0.5 e_{t-1} + 0.5 \frac{|S_t - S_{t-1}|}{S_{t-1}}, \tag{29}$$

where t is the iteration counter and S_t is the error at iteration t. The algorithm stops when e_t becomes less than some given threshold ϵ.

Let $\widehat{M}^* = (\widehat{m}_1^*, \ldots, \widehat{m}_n^*)$ denote the normalized credal partition obtained after convergence of the algorithm. The last step is to "denormalize" it using the masses $\widehat{m}(\emptyset)$ on the empty set computed in (15). This is done by multiplying each mass $\widehat{m}^*(A)$ with $A \neq \emptyset$ by $1 - \widehat{m}(\emptyset)$:

$$\widehat{m}_i(A) = [1 - \widehat{m}_i(\emptyset)] \, \widehat{m}_i^*(A), \tag{30}$$

for all $A \in 2^\Omega \setminus \{\emptyset\}$. The final consensus credal partition is $\widehat{M} = (\widehat{m}_1, \ldots, \widehat{m}_n)$. The whole procedure is summarized in Algorithm 1.

Concerning the time complexity of the method, converting each base credal partition to its relational representations requires $O(n^2)$ operations, where n is the number of objects, and combining the average partition takes $O(n^2 N)$. To compute the transitive closure, Lee [21] describes an optimal algorithm with $O(n^2)$ time complexity. Consequently, the calculation of the transitivized combined relational representation $\overline{\mathcal{R}}$ can be performed in $O(n^2 N)$ operations. The most computationally demanding step of the method is to recover the credal partition from the obtained relational representation. The complexity of this step depends on the Quadratic Programming (QP) problem (27) solved at each iteration. As the Jaccard matrix (11) is positive definite [5], so is matrix $\boldsymbol{\Sigma}$ in (27) (of size $f \times f$, where f is the number of focal sets). Consequently, the quadratic function (27) being minimized in convex. It is known [36] that convex QP problems can be solved in polynomial time. The computing time of the optimization can be controlled by limited the number of focal sets in the recovered credal partition.

Algorithm 1 Summary of the method.

Require: N credal partitions $M^b = (m_1^b, \ldots, m_n^b)^T$, $b = 1, ..., N$, combination rule, number k_r of clusters and focal sets $F_1, \ldots, F_f$ of the combined credal partition
Compute $\widehat{m}_i(\emptyset)$, $i = 1, \ldots, n$ using (15)
Normalize the base credal partitions by (16)
Compute the base relational representations $\mathcal{R}^{b*} = (m_{ij}^{b*})$, $b = 1, \ldots, N$
Compute the combined relational representation $\mathcal{R}^* = (m_{ij}^*)$
Compute the transitivized relational representation $\overline{R} = (\overline{m}_{ij})$
$t \leftarrow 0$, $e_0 \leftarrow 1$
Initialize $\mathbf{M}^*$ randomly, compute its relational representation
Compute S_0 using (19)
while $e_t \geq \epsilon$ **do**
  $t \leftarrow t + 1$
  $S_t \leftarrow 0$
  for $i = 1$ **to** n **do**
    Compute $\mathbf{A}_j$ for all $j \neq i$ using (24)
    Compute $\mathbf{\Sigma}$, $\mathbf{u}$ and c using (28)
    Find $\mathbf{m}_i^{*(t)}$ by minimizing (27) subject to $(\mathbf{m}_i^*)^T\mathbf{1} = 1$ and $\mathbf{m}_i^* \geq \mathbf{0}$
    Replace row i of $\mathbf{M}^*$ by $(\mathbf{m}_i^{*(t)})^T$
    $S_t \leftarrow S_t + g_i(\mathbf{m}_i^{*(t)})$
  end for
  $e_t \leftarrow 0.5e_{t-1} + 0.5|S_t - S_{t-1}|/S_{t-1}$
end while
Let $\widehat{M^*} = (\widehat{m}_1^*, \ldots, \widehat{m}_n^*)$ with $\widehat{m}_i^* = m_i^{*(t)}$, $i = 1, \ldots, n$
Denormalize the combined credal partition using (30)
return Credal partition $\widehat{M}$

4. Experiments

In this section, we evaluate our approach using various datasets. In Section 4.1, we study the influence of different parameters using simulated datasets. Detailed results with simulated and real datasets are then reported in Section 4.2.

4.1. Qualitative experiments

In this section, we present the results on simulated datasets to investigate different stages of the ensemble procedure. We study the effect of different combination rules and transitive closure with different t-norms.

Effect of different combination rules

We first investigate the effect of different combination rules on the Fourclass data, consisting of four classes in two-dimensional space, each generated from a Gaussian distribution. From Figure 1a, we can see that the clusters overlap, and there is an outlier.

In the first step, we ran the ECM algorithm [26, 6] with the Fourclass dataset for generating base partitions. To obtain useful information, we use the two-step approach introduced in [14, 34] to extract informative focal sets. More precisely, the ECM algorithm was first run with $\emptyset$, the singletons and Ω as focal sets. Based on the obtained credal partition, the similarity between each pairs of clusters $\{\omega_j, \omega_l\}$ was measured by $s(j,l) = \sum_{i=1}^{n} pl_{ij} pl_{il}$, where pl_{ij} and pl_{il} are the normalized plausibility that object i belongs, respectively, to cluster j and l. The pairs $\{\omega_j, \omega_l\}$ of mutual $K = 2$ nearest neighbors were then selected as informative focal sets. In the second step, the evidential clustering initialized with the credal partition computed in the previous step was run again with those informative focal sets. In [14], this method has been shown to be tractable and to yield good results even when the number of clusters is large.

In this experiment, we generated $N = 20$ base partitions, each of which had $k_b = 15$ clusters, and we set δ to the 0.2-quantile of the dissimilarities between objects. The diversity of the partitions is obtained by setting the number of clusters to a large number and using different random initializations. We considered the four combination rules reviewed in Section 2.1: average, Dempster, disjunctive and DP. For simplicity, after combining the relational representations we did not make the corresponding IFR transitive. We recovered a credal partition with four clusters, $k_r = 4$. In the second step, we firstly recovered a credal partition with the mass on the singletons and Ω, and secondly with the mass on the informative focal sets as we did in the generation step. One of the base partitions is shown in Figure 1b. In Figures 1b-1f, each point is represented by a symbol corresponding to its true class, and a color corresponding to the maximum plausibility cluster (6). The convex hulls of the lower and upper approximations computed using the interval dominance rule (7) are displayed using solid and broken lines, respectively.

From Figure 1d, we can see that Dempster's rule fails to capture the structure of this dataset. As recalled in Section 2.1, important assumptions for Dempster's rule are that all the mass functions come from reliable and independent sources; it is clear that none of these assumptions is satisfied in our case. When generating the base partitions, the number of clusters is set to a larger value than the number of "natural" clusters, and the base partitions, being based on the same data, are not independent. The DP rule behaves better than Dempster's rule, as it relaxed the assumption of perfect sources. However, the DP rule also cannot recognize the structure of the dataset (Figure 1f). The average and disjunctive rules both yield acceptable results (Figure 1c and 1e), but the credal partition obtained by the disjunctive rule is too imprecise: the inner approximations of the

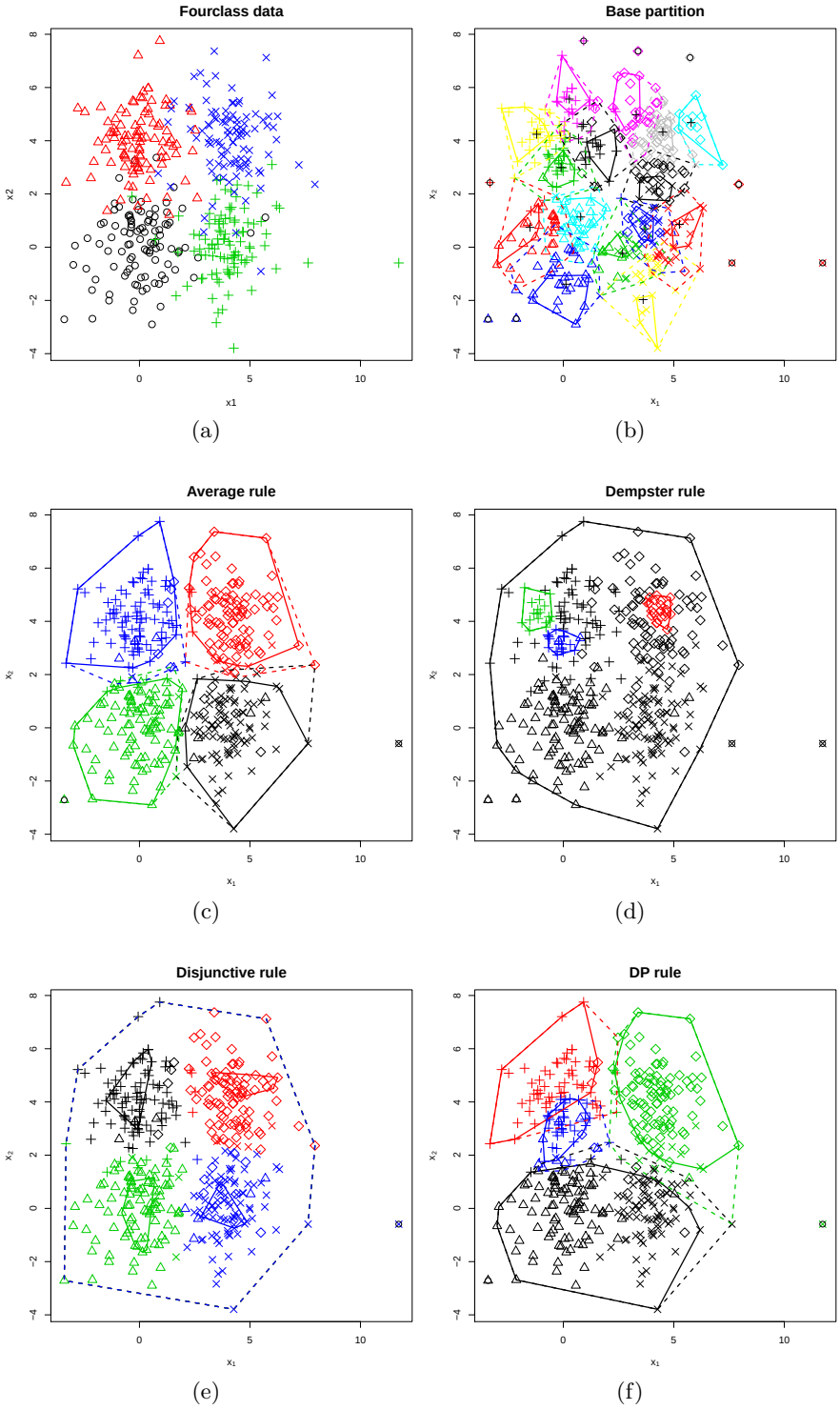

Figure 1: Results for the Fourclass data with different combination rules: original data (a), one of the base partitions (b), average rule (c), Dempster's rule (d), disjunctive rule (e), DP rule (f).

clusters are very small and the upper approximations are identical, which means that all objects possibly belong to any cluster (Figure 1e). This result is due to the very cautious assumption underlying this combination rule (at least one source is reliable). In contrast, the average rule successfully identifies the objects that can be clustered with high confidence (those in the inner approximations), as well as the objects in the overlapping area between clusters (Figure 1c). As the average rule appears to be the most effective, it was used in subsequent experiments.

Effect of transitive closure with different t-norms

In this section, we investigate the effect of transitive closure with respect to different t-norms on Half-rings data. This dataset is composed of two clusters in two-dimensional space, separated by a nonlinear boundary (Figure 2a). Such non-spherical clusters are typically difficult to identify without prior information. The parameter settings were the same as before, except that we set $k_b = 15$, $N = 20$, $k_r = 2$ and δ equal to the 0.1-quantile of the dissimilarities between objects in this experiment. We combined the relational representations by the average rule and made the belief and plausibility matrix transitive with respect to the minimum, Lukasiewicz and product t-norms. We also considered the results without transitivity comparison. One of the base partitions is shown in Figure 2b.

The recovered partitions with the different transitive closure operations are shown in Figure 2c-2e, and the recovered partition without transitivity is displayed in Figure 2f. In these figures, each point is represented by a symbol corresponding to its true class, and a color corresponding to the obtained cluster using the maximal plausibility rule. We can see that, without transitivity, our method fails to identify nonspherical clusters (Figure 2f). The results with the Lukasiewicz t-norm are similar to those without transitivity (Figure 2d). The best result is obtained with the minimum t-norm (Figure 2c).

Discovering clusters with complex shape is one of the most challenging issues in clustering. In this experiment, we have show that our method has the ability to discover such clusters after making the combined relational representation transitive in the sense of IFRs.

4.2. Quantitative experiments

In this section, we apply our method to the simulated and real data summarized in Table 1. Five simulated datasets[1] are shown in Figure 3; the first four datasets contain complex shape clusters, while R15 contains a comparatively larger number of well-separated clusters. All the real datasets can be found in the UCI Repository of machine learning databases[2]. For all the datasets considered in this study, we assume that the "ground-truth" partition exists. To compare an evidential partition with the true partition, we first converted it to a hard partition using the maximum plausibility rule, and we computed the adjusted Rand index (ARI) between the derived hard partition and the true partition.

We used the ECM algorithm with informative pair of clusters to generate base partitions for each dataset. We considered three cases for the number of clusters in the base partitions (Table 2): fixed small number (case 1); randomly selected from an interval (case 2) and fixed large number (case 3). We averaged the relational representations, and we computed the transitive closure operations with the minimum, product and Lukasiewicz t-norms. We also included the solution

[1] Available at `http://cs.joensuu.fi/sipu/datasets/`.
[2] Available at `http://archive.ics.uci.edu/ml`.

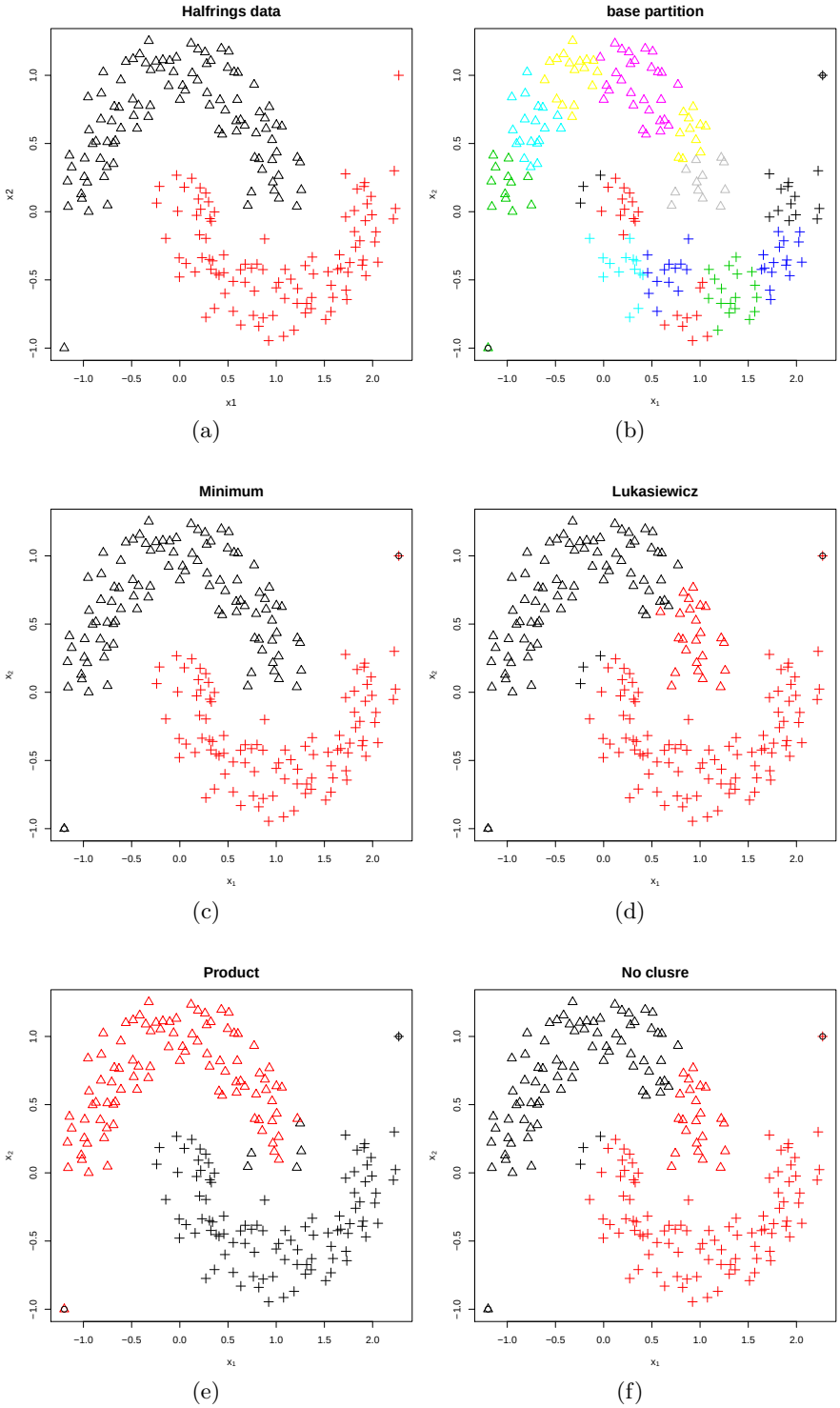

Figure 2: Results for Halfrings data: original data (a), one of base partitions (b), and recovered partition from transitive closures the minimum (c), Lukasiewicz (d) and product (e) t-norms, as well as without transitivity (f).

Table 1: Datasets used in the experiments.

Dataset	Number of objects	Number of clusters	Number of attributes
Aggregation	788	7	2
Compound	399	6	2
Flame	240	2	2
Spiral	312	3	2
R15	600	15	2
Iris	150	3	4
Seeds	210	3	7
Wine	178	3	13
Ecoli	307	4	5

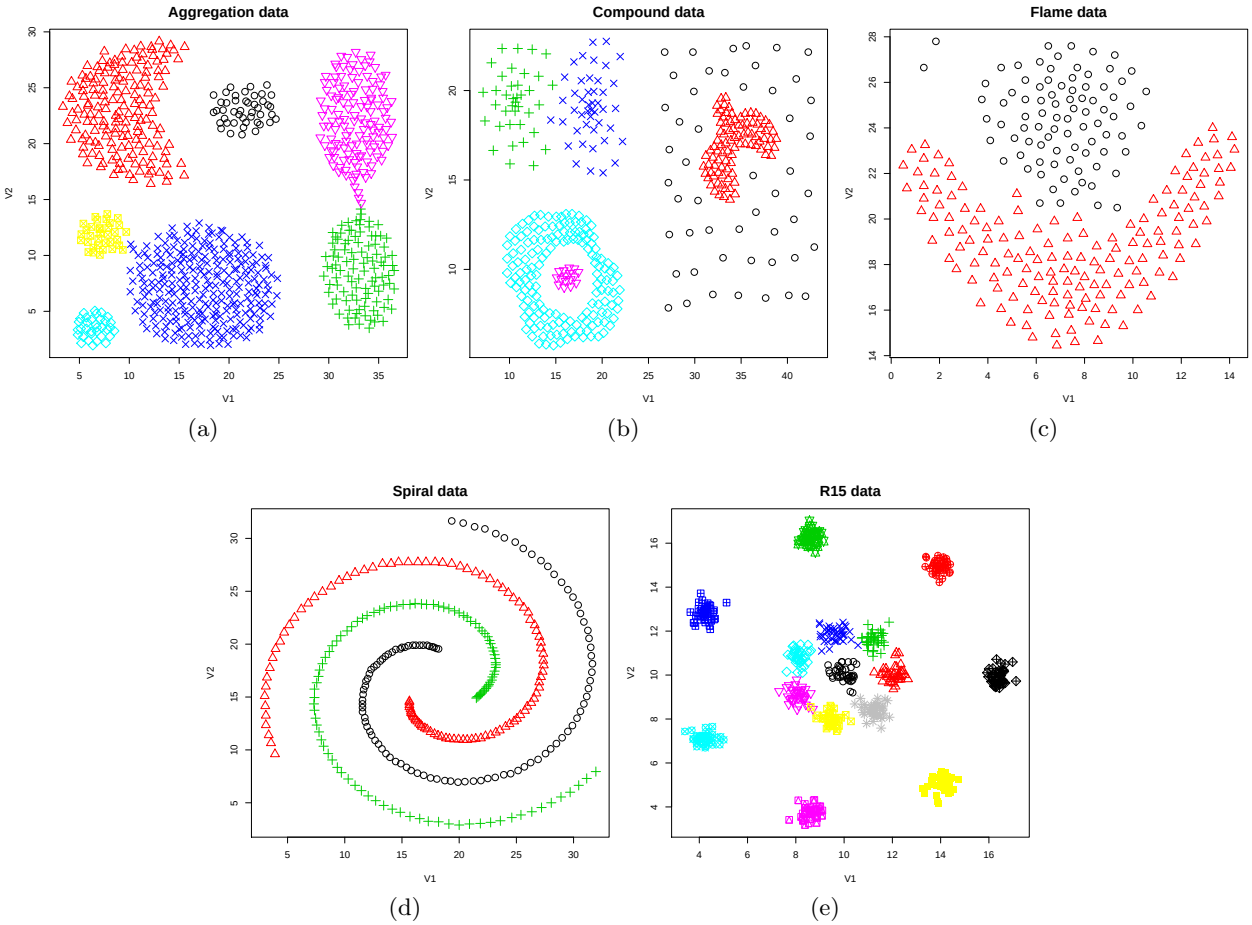

Figure 3: Simulated datasets: Aggregation (a), Compound (b), Flame (c), Spiral (d) and R15 (e).

Table 2: Number of clusters in base partitions

Dataset	Case 1	Case 2	Case 3
Aggregation	10	[10,20]	20
Compound	10	[10,20]	20
Flame	8	[8,15]	15
Spiral	30	[30,40]	40
R15	20	[20,30]	30
Iris	8	[8,15]	15
Seeds	5	[5,10]	10
Wine	5	[5,10]	10
Ecoli	8	[8,15]	15

without transitivity for comparison. The true number of clusters was assumed to be known; in practice, this number can often be guessed using, e.g., visualization techniques. When recovering the combined credal partition, we first used only singletons and Ω as focal sets (denoted by "simple" in the table), and then we included informative pairs informative pairs (denoted by "pairs" in the table) as explained in Section 4.1. The ensemble size was $N = 20$. The procedure was run 10 times for each experiment. The average ARI values are shown in Table 3, and the standard deviations are shown in parentheses.

For comparison, we show the results of the ECM algorithm with the true number of clusters. For this method, only $\emptyset$ and singletons were treated as focal sets, and we fixed $\delta = 100$. We also compared our method with the EAC method: in the first step we used the hard c-means algorithm to obtain $N = 100$ base partitions; in the second step, we used single-linkage hierarchical clustering to obtain the true number of clusters. For the EAC method, we also considered three cases as we did in our method. The results for the ECM algorithm and the EAC method are shown in the first and second columns of Table 3.

From Table 3, we can see that the results obtained by our method are better than those obtained by the ECM algorithm, except for the Wine data. When compared to the EAC method, our method achieves higher accuracy and better stability, especially for real datasets. The EAC method performs very well with simulated datasets, but better or similar results were obtained with our method, except for the Spiral dataset.

Comparing the results obtained with different t-norms for the transitive closure operation, we can see that the minimum t-norm often performs well, except with the R15 and Seeds dataset, for which better results are obtained with the other t-norms, or even without transitivity. The results with the Lukasiewicz and product t-norms are often similar to those obtained without transitivity. Generally, it seems that making the combined relational minimum-transitive is beneficial for datasets with complex-shaped clusters, but it can sometimes degrade the performance for datasets with overlapping clusters. The reason is that, where there is an overlapping area between clusters and we make the fuzzy relation transitive, objects from different clusters become similar to each other, which hinders the performance of the method. As far as the number of clusters in the base partitions is concerned, better results are generally obtained in Cases 2 and 3, i.e., with a larger number of clusters.

Table 3: Average ARI results for simulated and real datasets. The best results are shown in bold.

		ECM	EAC	minimum		Lukasiewicz		product		no transitivity	
				simple	pairs	simple	pairs	simple	pairs	simple	pairs
Aggregation	case 1	0.68(0.05)	0.81(0.06)	**0.83(0.08)**	0.73(0.08)	0.72(0.07)	0.72(0.07)	0.79(0.07)	0.82(0.09)	0.72(0.06)	0.71(0.05)
	case 2		0.87(0.04)	**0.97(0.04)**	0.95(0.07)	0.69(0.05)	0.67(0.05)	0.82(0.07)	0.76(0.05)	0.76(0.06)	0.7(0.05)
	case 3		0.91(0.07)	**0.98(0.01)**	**0.98(0.01)**	0.73(0.04)	0.71(0.05)	0.81(0.08)	0.77(0.06)	0.74(0.03)	0.73(0.04)
Compound	case 1	0.53(0.09)	0.65(0.09)	**0.78(0.02)**	0.72(0.03)	0.52(0.01)	0.52(0.01)	0.51(0.04)	0.5(0.06)	0.52(0.04)	0.52(0.04)
	case 2		0.86(0.03)	**0.87(0.04)**	0.78(0.07)	0.48(0.06)	0.49(0.05)	0.5(0.04)	0.47(0.06)	0.49(0.05)	0.49(0.05)
	case 3		**0.89(0.01)**	0.73(0.1)	**0.88(0.05)**	0.41(0.02)	0.41(0.02)	0.45(0.06)	0.45(0.06)	0.42(0.04)	0.39(0)
Flame	case 1	0.49(0)	0.93(0)	0.9(0.02)	0.9(0.02)	0.75(0.14)	0.75(0.14)	**0.92(0.01)**	**0.92(0.01)**	0.79(0.05)	0.79(0.05)
	case 2		0.19(0.39)	0.91(0.02)	0.91(0.02)	0.69(0.33)	0.69(0.33)	**0.93(0.04)**	**0.93(0.04)**	0.82(0.03)	0.82(0.03)
	case 3		-0.01(0.06)	**0.92(0.01)**	**0.92(0.01)**	0.47(0.25)	0.47(0.25)	0.66(0.23)	0.66(0.23)	0.53(0.3)	0.53(0.3)
Spiral	case 1	0.01(0)	**1(0)**	0.82(0.21)	0.72(0.28)	0.01(0)	0.01(0)	0.01(0.01)	0.01(0)	0.01(0)	0.01(0)
	case 2		**1(0)**	0.88(0.16)	0.87(0.16)	0.03(0.01)	0.03(0.01)	0.05(0.03)	0.03(0.01)	0.04(0.04)	0.04(0.01)
	case 3		**1(0)**	0.96(0.06)	0.76(0.19)	0.08(0.06)	0.11(0.07)	0.14(0.02)	0.14(0.11)	0.06(0.04)	0.03(0.01)
R15	case 1	0.93(0.07)	**0.99(0)**	0.56(0.12)	0.57(0.08)	**0.99(0)**	**0.99(0)**	0.98(0.03)	0.97(0.03)	**0.99(0)**	**0.99(0)**
	case 2		**0.99(0)**	0.67(0.09)	0.62(0.07)	**0.99(0)**	**0.99(0)**	**0.99(0)**	**0.99(0)**	**0.99(0)**	**0.99(0)**
	case 3		**0.99(0)**	0.86(0.02)	0.66(0.11)	**0.99(0)**	**0.99(0)**	**0.99(0)**	**0.99(0)**	**0.99(0)**	**0.99(0)**
Iris	case 1	0.73(0)	0.6(0.04)	0.7(0.03)	0.69(0.05)	0.84(0.03)	0.8(0.04)	**0.86(0.03)**	0.85(0.03)	0.84(0.03)	0.8(0.03)
	case 2		0.57(0)	0.76(0.06)	0.68(0.04)	0.79(0.05)	0.73(0.02)	0.8(0.05)	0.75(0.03)	**0.82(0.05)**	0.74(0.02)
	case 3		0.57(0)	**0.85(0.04)**	0.78(0.04)	0.77(0.05)	0.7(0.03)	0.76(0.03)	0.7(0.03)	0.74(0.11)	0.69(0.03)
Seeds	case 1	0.78(0.01)	0.29(0.09)	0.48(0)	0.48(0.02)	0.78(0.01)	0.78(0.02)	0.78(0)	**0.79(0)**	0.78(0.01)	0.78(0.01)
	case 2		0.22(0.08)	0.72(0.03)	0.66(0.08)	0.77(0.02)	**0.78(0.02)**	0.75(0.01)	0.75(0.02)	**0.78(0.02)**	0.77(0.02)
	case 3		0.24(0)	0.73(0.09)	0.55(0.07)	**0.77(0.02)**	0.72(0.04)	0.74(0.03)	0.7(0.04)	0.76(0.04)	0.7(0.06)
Wine	case 1	**0.9(0)**	0.35(0.24)	0.66(0.18)	0.8(0.09)	0.71(0.13)	**0.88(0.02)**	0.78(0.18)	0.85(0.07)	0.77(0.16)	0.87(0.01)
	case 2		0.09(0.2)	**0.78(0.09)**	0.73(0.13)	0.68(0.17)	0.7(0.17)	0.73(0.18)	0.73(0.21)	0.7(0.16)	0.69(0.13)
	case 3		0.17(0.23)	**0.83(0.05)**	**0.83(0.05)**	0.58(0.17)	0.58(0.17)	0.78(0.1)	0.8(0.07)	0.71(0.19)	0.72(0.19)
Ecoli	case 1	0.52(0)	0.47(0.16)	0.57(0.07)	0.62(0.1)	0.64(0.09)	0.62(0.09)	**0.65(0.09)**	0.6(0.09)	0.62(0.09)	0.62(0.09)
	case 2		0.51(0.04)	0.6(0.08)	0.55(0.11)	0.62(0.1)	0.63(0.1)	0.66(0.08)	0.67(0.07)	**0.69(0.06)**	0.69(0.07)
	case 3		0.48(0.01)	0.45(0.11)	0.47(0.13)	0.62(0.13)	**0.63(0.11)**	0.62(0.14)	0.62(0.1)	0.62(0.1)	0.61(0.1)

5. Conclusion

We have presented a method for combining clusterings in the DS framework. Each base clustering is assumed to take the form of a credal partition, in which the clustering membership of each object is allowed to be uncertain and represented by a mass function. This very general formalism encompasses hard, fuzzy and rough partitions as special cases [9]. Credal partitions of special forms can be generated by hard, fuzzy or rough clustering algorithms, and general credal partitions can be obtained by evidential clustering procedures such a EVCLUS [14], ECM [26], BPEC [34], CCM [25], etc.

Base credal partitions cannot be combined directly, because there is not always a clear correspondence between clusters in different partitions; in particular, base credal partitions can have different numbers of clusters. To circumvent this difficulty, we proposed to convert each base credal partition to its relational representation, defined as the collection of pairwise mass functions describing the uncertain joint cluster-membership for each pair of objects. After the normalized relational representations have been computed, they can be aggregated using any combination rule of DS theory. The best results have been obtained with the averaging operator.

Using the similarity between relational representations and intuitionistic fuzzy relations studied in [22], we have proposed a way to transitivize the combined relational representation by computing the transitive closures of two fuzzy relations, based on a t-norm. Our experimental results suggest that the minimum t-norm often yields the best results, especially for datasets with complex-shaped clusters. However, making the relational representation transitive does not always improve the results, and may even degrade them in the case of datasets with many spherical clusters.

After the combined relational representation has been computed, the last step of our method consists in constructing a credal partition whose relational representation is as close as possible to the combined relational representation obtained in the previous step. We have proposed an error measure based on Jousselme's metric, which can be minimized using a grouped coordinate descent algorithm that solves a convex quadratic optimization problem at each step. After a normalized credal partition has been obtained, we "denormalize" it by assigning to the empty set the average of the masses assigned to the empty set by the base partitions, which provide useful information to signal outliers.

We have applied this method to a variety of simulated and real datasets. It has been shown to perform well in terms of adjusted Rand index as compared to the EAC method and to the ECM algorithm alone. It should also be emphasized that, in contrast with EAC and most existing ensemble clustering methods, our approach computes a credal partition, which constitutes a richer description of the clustering structure of a dataset, as compared to hard or fuzzy partitions.

Although very encouraging, these results are still preliminary. The determination of the number of clusters remains a crucial issue that remains to be thoroughly investigated. The application of this approach to very big datasets with a large number of clusters still represents a challenge. Finally, we could apply this approach not only to combine credal partitions, but also to combine all kinds of partitions generated by all kinds of clustering algorithms. These research directions will be investigated in future work.

References

[1] Akbari, E., Dahlan, H. M., Ibrahim, R., Alizadeh, H., 2015. Hierarchical cluster ensemble selection. Engineering Applications Of Artificial Intelligence 39, 146–156.

[2] Antoine, V., Quost, B., Masson, M.-H., Denoeux, T., 2012. CECM: Constrained evidential c-means algorithm. Computational Statistics & Data Analysis 56 (4), 894–914.
[3] Antoine, V., Quost, B., Masson, M.-H., Denoeux, T., 2014. CEVCLUS: Evidential clustering with instance-level constraints for relational data. Soft Computing 18 (7), 1321–1335.
[4] Boixader, D., Jacas, J., Recasens, J., 2000. Fuzzy equivalence relations: advanced material. In: Dubois, D., Prade, H. (Eds.), Fundamentals of Fuzzy sets. Kluwer Academic Publishers, Boston, pp. 261–290.
[5] Bouchard, M., Jousselme, A.-L., Doré, P.-E., 2013. A proof for the positive definiteness of the Jaccard index matrix. International Journal of Approximate Reasoning 54 (5), 615–626.
[6] Denœux, T., 2016. evclust: Evidential Clustering. R package version 1.0.3.
URL `https://CRAN.R-project.org/package=evclust`
[7] Denoeux, T., 2019. Decision-making with belief functions: a review. International Journal of Approximate Reasoning 109, 87–110.
[8] Denœux, T., Dubois, D., Prade, H., 2020. Representations of uncertainty in artificial intelligence: Beyond probability and possibility. In: Marquis, P., Papini, O., Prade, H. (Eds.), A Guided Tour of Artificial Intelligence Research. Springer Verlag, Ch. 4.
[9] Denoeux, T., Kanjanatarakul, O., September 2016. Beyond fuzzy, possibilistic and rough: An investigation of belief functions in clustering. In: Soft Methods for Data Science (Proc. of the 8th International Conference on Soft Methods in Probability and Statistics SMPS 2016). Vol. AISC 456 of Advances in Intelligent and Soft Computing. Springer-Verlag, Rome, Italy, pp. 157–164.
[10] Denoeux, T., Kanjanatarakul, O., Sriboonchitta, S., 2015. EK-NNclus: a clustering procedure based on the evidential k-nearest neighbor rule. Knowledge-Based Systems 88, 57–69.
[11] Denoeux, T., Li, S., Sriboonchitta, S., 2018. Evaluating and comparing soft partitions: An approach based on Dempster-Shafer theory. IEEE Transactions on Fuzzy Systems 26 (3), 1231–1244.
[12] Denœux, T., Masson, M.-H., 2004. EVCLUS: Evidential clustering of proximity data. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 34 (1), 95–109.
[13] Denœux, T., Masson, M.-H., 2011. Evidential reasoning in large partially ordered sets. Annals of Operations Research 195 (1), 135–161.
[14] Denœux, T., Sriboonchitta, S., Kanjanatarakul, O., 2016. Evidential clustering of large dissimilarity data. Knowledge-Based Systems 106, 179–195.
[15] Dubois, D., Prade, H., 1988. Representation and combination of uncertainty with belief functions and possibility measures. Computational Intelligence 4 (3), 244–264.
[16] Fred, A. L. N., Jain, A. K., 2002. Data clustering using evidence accumulation. In: International Conference on Pattern Recognition. pp. 276–280.
[17] Fred, A. L. N., Jain, A. K., 2005. Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (6), 835–850.
[18] Ghaemi, R., Sulaiman, m. n., Ibrahim, H., Mustapha, N., 2009. A survey: Clustering ensembles techniques. World Academy of Science, Engineering and Technology 38, 644–653.
[19] Iam-on, N., Boongoen, T., Garrett, S., 2008. Refining pairwise similarity matrix for cluster ensemble problem with cluster relations. Discovery Science, 222–233.
[20] Jousselme, A.-L., Grenier, D., Bossé, E., 2001. A new distance between two bodies of evidence. Information Fusion 2 (2), 91–101.
[21] Lee, H.-S., 2001. An optimal algorithm for computing the max-min transitive closure of a fuzzy similarity matrix. Fuzzy Sets and Systems 123 (1), 129–136.
[22] Li, B., He, W., 2014. The structures of intuitionistic fuzzy equivalence relations. Information Sciences 278, 883–899.
[23] Li, F., Li, S., Denoeux, T., 2018. k-CEVCLUS: Constrained evidential clustering of large dissimilarity data. Knowledge-Based Systems 142, 29–44.
[24] Li, F., Qian, Y., Wang, J., Liang, J., 2017. Multigranulation information fusion: A Dempster-Shafer evidence theory-based clustering ensemble method. Information Sciences 378, 389–409.
[25] Liu, Z., Pan, Q., Dezert, J., Mercier, G., 2015. Credal c-means clustering method based on belief functions. Knowledge-Based Systems 74, 119–132.
[26] Masson, M.-H., Denoeux, T., 2008. ECM: An evidential version of the fuzzy c-means algorithm. Pattern Recognition 41 (4), 1384–1397.
[27] Masson, M.-H., Denœux, T., 2009. RECM: Relational evidential c-means algorithm. Pattern Recognition Letters 30 (11), 1015–1026.
[28] Masson, M.-H., Denoeux, T., 2011. Ensemble clustering in the belief functions framework. International Journal

of Approximate Reasoning 52 (1), 92–109.
[29] Ovchinnikov, S., 2000. An introduction to fuzzy relations. In: Dubois, D., Prade, H. (Eds.), Fundamentals of Fuzzy sets. Kluwer Academic Publishers, Boston, pp. 233–259.
[30] Peters, G., Crespo, F., Lingras, P., Weber, R., 2013. Soft clustering: fuzzy and rough approaches and their extensions and derivatives. International Journal of Approximate Reasoning 54 (2), 307–322.
[31] Shafer, G., et al., 1976. A Mathematical Theory of Evidence. Princeton university press.
[32] Smets, P., 1990. The combination of evidence in the transferable belief model. IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (5), 447–458.
[33] Smets, P., 1993. Belief functions: The disjunctive rule of combination and the generalized bayesian theorem. International Journal of Approximate Reasoning 9 (1), 1–35.
[34] Su, Z.-G., Denoeux, T., 2019. BPEC: Belief-peaks evidential clustering. IEEE Transactions on Fuzzy Systems 27 (1), 111–123.
[35] ter Braak, C. J. F., Kourmpetis, Y., Kiers, H. A. L., Bink, M. C. A. M., 2009. Approximating a similarity matrix by a latent class model: a reappraisal of additive fuzzy clustering. Computational Statistics & Data Analysis 53 (8), 3183–3193.
[36] Vavasis, S. A., 2001. Complexity theory: quadratic programming. Springer US, Boston, MA, pp. 304–307.
[37] Vega-Pons, S., Ruiz-Shulcloper, J., 2009. Clustering ensemble method for heterogeneous partitions. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Vol. 5856. pp. 481–488.
[38] Vega-Pons, S., Ruiz-Shulcloper, J., 2011. A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence 25 (03), 337–372.
[39] Wang, X., Han, D., Han, C., 2017. Ensemble clustering based on evidence theory. In: 2017 20th International Conference on Information Fusion (Fusion). pp. 759–767.
[40] Wang, X., Yang, C., Zhou, J., 2009. Clustering aggregation by probability accumulation. Pattern Recognition 42 (5), 668–675.
[41] Yang, L., Lv, H., Wang, W., 2006. Soft cluster ensemble based on fuzzy similarity measure. In: IMACS: Multiconference on Computational Engineering in Systems Applications, Vols 1 and 2. pp. 1994 – 1997.
[42] Zhan, J.-M., Chen, J.-T., Xing, J.-Q., 2017. Research advance of clustering ensemble algorithm. In: 2017 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR). pp. 109–114.
[43] Zhou, K., Martin, A., Pan, Q., Liu, Z.-G., 2015. Median evidential c-means algorithm and its application to community detection. Knowledge-Based Systems 74, 69–88.