CECM : Constrained Evidential C-Means algorithm

V. Antoine[a,c], B. Quost[a,c], M.-H. Masson[b,c], T. Denœux[a,c]

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

Abstract

The aim of cluster analysis is to group objects according to their similarity. Some methods use hard partitioning, some use fuzzy partitioning and, recently, a new concept of partition based on belief function theory, called credal partition, has been proposed. It makes it possible to generate meaningful representations of the data and to improve robustness with respect to outliers. All these methods are unsupervised ones, as the similarity between the objects is determined using only a numeric description of the objects. However, in some applications, some kind of background knowledge about the objects or about the clusters is available. To integrate this auxiliary information, constraint-base (or semi-supervised) methods have been proposed. A popular type of constraints specifies whether two objects are in the same cluster (must-link) or in different clusters (cannot-link). Moreover, actively selecting object pairs allows us to get improved clustering performances using only a small number of constraints. We propose here a new algorithm, called CECM, which combines belief functions and the constrained clustering frameworks. We show how to translate the available information into constraints and how to integrate them in the search of a credal partition. The paper ends with some experimental results on synthetic and real data. In particular, we show how CECM may be used to integrate prior knowledge in a medical image segmentation task.

Keywords: Clustering, semi-supervised learning, pairwise constraints, adaptive metric, active learning; belief functions, Dempster-Shafer theory, evidence theory.

1. Introduction

Clustering is a classical data analysis method that aims at grouping a set of objects into clusters based on similarity between their descriptors. However, there are some situations in which some background knowledge about the problem is available. Making use of this extra-information in a clustering algorithm

Email address: violaine.antoine@hds.utc.fr (V. Antoine)

can help us to guide the method towards a desired solution and to improve the classification accuracy. Prior information can be exploited at different levels of the classification such as: the *cluster* level with, for instance, a minimum distance neighbourhood [8], the *model* level with the requirement of balanced clusters [31] or the specification of non desired solutions [11], or at the *instance* level. Wagstaff [27] proposed to introduce two types of instance-level constraints: the first one specifies that two objects have to be in the same cluster (*must-link* constraint) while the second one specifies that two objects should not be put in the same cluster (*cannot-link* constraint). Such pairwise constraints have been considered and integrated in many unsupervised algorithms such as the hard or the fuzzy c-means (FCM), and have recently become a topic of great interest [28, 3, 26, 8, 19]. They have been incorporated in many different ways, generally by including a penalty term in the objective function [2, 12] or by altering the distances between objects with respect to the constraints [15, 28].

In the FCM algorithm, each object may belong to one or more clusters with different degrees of membership. These degrees of membership are stored into a fuzzy partition matrix $U = (u_{ik})$ and are calculated by minimizing a suitable objective function with respect to the constraints

$$u_{ik} \geq 0 \quad \forall\, i, k, \tag{1}$$

$$\sum_{k} u_{ik} = 1. \tag{2}$$

Each number $u_{ik} \in [0, 1]$ is interpreted as the degree of membership of object i to cluster k. Nevertheless the method sometimes produces counterintuitive results and has poor robustness against noise and outliers. That is the reason why possibilistic methods [17, 22] and, more recently, algorithms using the theoretical framework of belief functions [10, 20, 21] have been proposed. All these works are based on a new concept of partition, referred to as a *credal* partition, which extends the existing concepts of hard, fuzzy and possibilistic partitions. This is done by allocating, for each object, a mass of belief, not only to single clusters, but also to any subset of the set of clusters $\Omega = \{\omega_1, \ldots, \omega_c\}$. Experiments have shown that this additional flexibility allows us to gain a deeper insight into the data and to improve robustness with respect to outliers. One of the algorithms designed to derive a credal partition from data, called Evidential C-Means (ECM), can be considered as a direct extension of FCM. In this paper, we propose to introduce pairwise constraints in the ECM algorithm, in order to create a new algorithm, called CECM, which will combine the advantages of adding background knowledge and using belief functions. Furthermore, we present a formulation of ECM that adapts the metric using a Mahalanobis distance so that the constraints may be more easily satisfied. Finally we propose an active learning scheme, based on the credal partition, which makes it possible to select efficient pairwise constraints.

The remaining of this paper is organized as follows. First, a brief overview of the theory of belief functions is provided in Section 2. In the same section, the main fuzzy partitioning algorithms from which ECM is derived are

presented. Then the notion of credal partition and the way to derive it from data are described. Some useful interpretation tools are also recalled. Section 3 introduces the algorithm CECM. First, we show how to translate in a natural way the available information in terms of constraints on belief masses. Then we explain how to integrate these constraints in the search of the credal partition. In Section 4, we describe a version of CECM with an adaptive metric. The last part describes the experimental settings and the results. Several results are presented, considering that the constraints are either available *a priori* or gradually selected during the learning phase. Finally, some perspectives of our work are presented in a conclusion.

2. Background

2.1. Belief functions

The Dempster-Shafer theory of evidence [23, 25] (or belief function theory) is a theoretical framework for representing partial and unreliable information. In this section, only the main concepts are recalled.

Let us consider a variable ω taking values in a finite set $\Omega = \{\omega_1, \ldots, \omega_c\}$ called the frame of discernment. Partial knowledge regarding the actual value taken by ω can be represented by a mass function m, which is an application from the power set of Ω in the interval $[0, 1]$ such that

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{3}$$

The subsets A of Ω such that $m(A) > 0$ are called focal sets of m. The value of the focal set $m(A)$ can be interpreted as a fraction of a unit mass of belief that is allocated to A and that cannot be allocated to any subset of A. Complete ignorance is obtained when Ω is the only focal set, and full certainty when the whole mass of belief is assigned to a unique singleton of Ω (m is then said to be a *certain* bba). If all the focal sets of m are singletons, m is similar to a probability distribution: it is then called a *Bayesian* bba. A bba m such that $m(\emptyset) = 0$ is said to be normalized. Under the *open-world* assumption, a mass function $m(\emptyset) > 0$ is interpreted as a quantity of belief given to the hypothesis that the actual value of ω might not belong to Ω [24].

Given a mass function m, it is possible to define a plausibility function $pl : 2^\Omega \rightarrow [0, 1]$ and a belief function $bel : 2^\Omega \rightarrow [0, 1]$ by:

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

$$bel(A) = \sum_{B \subseteq A, B \neq \emptyset} m(B) \quad \forall A \subseteq \Omega. \tag{5}$$

Functions bel and pl are linked by the following relation:

$$pl(A) = 1 - m(\emptyset) - bel(\overline{A}), \tag{6}$$

where $\overline{A}$ denotes the complement of A. The quantity $bel(A)$ is interpreted as a degree of belief in A, taking into account the mass of belief given to A and nonempty subsets of A. In contrast, $pl(A)$ measures to what extent one fails to believe in $\overline{A}$.

In order to make a decision regarding the value of ω, it is possible to transform the mass function into a pignistic probability distribution [25], defined, for a normal bba, as:

$$BetP(\omega) = \sum_{\omega \in A} \frac{m(A)}{|A|} \quad \forall \omega \in \Omega, \tag{7}$$

where $|A|$ denotes the cardinality of $A \subseteq \Omega$. If $m(\emptyset) \neq 0$, then a normalization step has to be performed before carrying out the pignistic transformation. Various methods may be applied. In particular, Dempster's normalization consists in dividing all the masses by $1 - m(\emptyset)$, whereas Yager's normalization transfers $m(\emptyset)$ to $m(\Omega)$ [29].

2.2. Fuzzy c-means and variants

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a collection of vectors in $\mathbb{R}^p$ describing n objects to classify in the set $\Omega = \{\omega_1 \ldots \omega_c\}$. Each cluster ω_k, $k = 1, c$ is represented by a prototype or a centroid $\mathbf{v}_k \in \mathbb{R}^p$. Let V denote the matrix composed of the cluster centroids, and let $U = (u_{ik})$ define a fuzzy partition matrix that contains the degrees of membership of each object to each cluster. The FCM algorithm [4] computes V and U so as to minimize the following objective function:

$$J_{\text{FCM}}(U, V) = \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik}^{\beta} d_{ik}^2, \tag{8}$$

subject to (1) et (2). In the objective function (8), d_{ik} represents the Euclidean distance between the object $\mathbf{x}_i$ and the centroid $\mathbf{v}_k$ and $\beta > 1$ is a weighting exponent that controls the fuzziness of the partition. The objective function is minimized using an iterative algorithm, which alternatively optimizes the cluster centers and the membership degrees. The update formulas for the masses and the centers are obtained by computing the Lagrangian formulation of the optimization problem and by setting its partial derivatives with respect to the parameters to zero [4]. We obtain:

$$\mathbf{v}_k = \frac{\sum_{i=1}^{n} u_{ik}^{\beta} \mathbf{x}_i}{\sum_{i=1}^{n} u_{ik}^{\beta}} \quad k = 1, c, \tag{9}$$

$$u_{ij} = \frac{d_{ij}^{-2/(\beta-1)}}{\sum_{k=1}^{c} d_{ik}^{-2/(\beta-1)}} \quad i = 1, n \quad j = 1, c. \tag{10}$$

The algorithm starts from an initial guess for either the partitioning matrix or the cluster centers and iterates until convergence.

To detect noisy data or outliers, Davé [7] has proposed a variant of FCM called the "noise-clustering" algorithm (NC). It consists in adding to the c initial

clusters a "noise" cluster, associated to a fixed distance ρ to all objects. The parameter ρ controls the amount of data considered as outliers. The membership u_{i*} of an object i to the noise cluster is given by:

$$u_{i*} = 1 - \sum_{k=1}^{c} u_{ik} \quad i = 1, n, \tag{11}$$

The objective function to be minimized can be written as:

$$J_{\text{NC}}(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{\beta} d_{ij}^2 + \sum_{i=n}^{c} \rho^2 u_{i*}^{\beta}. \tag{12}$$

Writing the optimality conditions of the problem leads, as in FCM, to direct adaptation formulas for the memberships and the cluster centers.

The Gustafson and Kessel algorithm [13] is another interesting variant of FCM. This algorithm extends FCM by using an adaptive distance, in order to detect clusters of different geometrical shapes. Each cluster has its own norm-inducing matrix S_k defined as its fuzzy covariance matrix:

$$S_k = \frac{\sum_{i=1}^{n} u_{ik}^{\beta} (\mathbf{x}_i - \mathbf{v}_k)(\mathbf{x}_i - \mathbf{v}_k)^t}{\sum_{i=1}^{n} u_{ik}^{\beta}} \quad k = 1, c \quad i = 1, n. \tag{13}$$

Then, the distance between an object $\mathbf{x}_i$ and a center $\mathbf{v}_k$ is taken as:

$$d_{ik}^2 = \det(S_k)^{\frac{1}{p}} (\mathbf{x}_i - \mathbf{v}_k)^t S_k^{-1} (\mathbf{x}_i - \mathbf{v}_k). \tag{14}$$

Equation (13) can be obtained by imposing a constant volume to each cluster and using Lagrange multipliers, except for the normalization by the factor $\sum_{i=1}^{n} u_{ik}^{\beta}$ (which could be omitted). Additionally, Gustafson and Kessel show that the adaptation formulas of FCM for the membership degrees and the centers remain valid as they do not depend on the metric.

2.3. ECM algorithm

Recently, Masson and Denœux proposed a credibilistic version of Davé's algorithm [20] by replacing the fuzzy partition matrix U with a more general kind of partition M called a credal partition. In this framework, partial knowledge regarding the class membership of an object is represented by a mass function on the set Ω of possible classes. Thus, belief mass may be given to any subset A of Ω (any set of classes), and not only to singletons of Ω. This representation enables to model a wide variety of situations ranging from complete ignorance to full certainty, as illustrated in Example 1.

Example 1. *Let us consider a collection of four objects that need to be classified into two classes. A credal partition is presented in Table 1. The class of the first object is known with certainty, whereas the class of the second object is completely unknown. We have probabilistic knowledge of the actual class of the third object. The last object is considered to be an outlier, which is represented by allocating the whole unit mass to the empty set.*

Table 1: Example of a credal partition

A	$m_1(A)$	$m_2(A)$	$m_3(A)$	$m_4(A)$
$\emptyset$	0	0	0	1
$\{\omega_1\}$	1	0	0.3	0
$\{\omega_2\}$	0	0	0.7	0
Ω	0	1	0	0

A credal partition can thus be seen as a general model of partitioning:

- when each m_i is a *certain* bba, then M defines a conventional, crisp partition of the set of objects; this corresponds to a situation of complete knowledge;
- when each m_i is a Bayesian bba, then M specifies a fuzzy partition;
- when the focal elements of all bbas are restricted to be singletons of Ω or the empty set, a partition with a noise cluster as in the NC algorithm is recovered.

ECM is one of the algorithms proposed to derive a credal partition from data. Let m_{ij} denote the degree of belief that object $\mathbf{x}_i$ belongs to the subset $A_j \subseteq \Omega$. Deriving a credal partition implies determining for each object $\mathbf{x}_i$ the quantities $m_{ij} = m_i(A_j)\ \forall A_j \neq \emptyset, A_j \subseteq \Omega$ in such a way that a low (resp., high) value of m_{ij} is found when the distance d_{ij} between $\mathbf{x}_i$ and A_j is high (resp., low). The distance d_{ij} between an object and a set of classes A_j is defined as follows. Like in fuzzy partitioning, each class ω_l is represented by a center $\mathbf{v}_l \in \mathbb{R}^p$. Then, for each subset $A_j \subseteq \Omega$, $A_j \neq \emptyset$, a centroid $\overline{\mathbf{v}}_j$ is calculated as the barycenter of the centers associated to the classes in A_j:

$$\overline{\mathbf{v}}_j = \frac{1}{|A_j|} \sum_{l=1}^{c} s_{lj} \mathbf{v}_l, \tag{15}$$

with

$$s_{lj} = \begin{cases} 1 & \text{if } \omega_l \in A_j, \\ 0 & \text{else.} \end{cases} \tag{16}$$

The distance d_{ij} between $\mathbf{x}_i$ and the focal set A_j may then be defined by:

$$d_{ij} = ||\mathbf{x}_i - \overline{\mathbf{v}}_j||. \tag{17}$$

The ECM algorithm searches for the M and V matrices that minimize a criterion similar to that of the NC algorithm:

$$J_{\text{ECM}}(M, V) = \sum_{i=1}^{n} \sum_{A_k \neq \emptyset} |A_k|^{\alpha} m_{ik}^{\beta} d_{ik}^2 + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta}, \tag{18}$$

subject to :

$$\sum_{k/A_k \subseteq \Omega, A_k \neq \emptyset} m_{ik} + m_{i\emptyset} = 1 \quad \forall i = 1, n, \tag{19}$$

where $m_{i\emptyset}$ denotes the mass of the object $\mathbf{x}_i$ allocated to the empty set. The empty set is interpreted as a noise cluster; thus, it is dealt with separately. The parameter ρ represents the distance of all the objects to the empty set. An additional weighting coefficient $|A_k|^\alpha$ is introduced to penalize the allocation of belief to subsets with high cardinality, the exponent α allowing us to control the degree of penalization.

As in FCM or NC, the credal partition is found by performing an iterative optimization with the alternate update of the masses and the centroids. The necessary condition of optimality for M gives the following adaptation rule for the mass functions:

$$m_{ij} = \frac{|A_j|^{-\alpha/(\beta-1)} d_{ij}^{-2/(\beta-1)}}{\sum_{A_k \neq \emptyset} |A_k|^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)} + \rho^{-2/(\beta-1)}} \quad i = 1, n \quad \forall A_j \neq \emptyset \tag{20}$$

and

$$m_{i\emptyset} = 1 - \sum_{A_j \neq \emptyset} m_{ij} \quad i = 1, n. \tag{21}$$

Note that these update equations are very similar to those of the NC algorithm except that there are 2^c values m_{ij} to compute instead of $c+1$ fuzzy membership degrees u_{ij}. A more complex update rule is found for the centroids, since the optimality conditions lead to the resolution of a linear system at each step of the optimization process. Let $\mathbf{B}$ be a matrix of size $(c \times p)$ defined by:

$$\mathbf{B}_{lq} = \sum_{i=1}^{n} x_{iq} \sum_{A_j \neq \emptyset} |A_j|^{\alpha-1} m_{ij}^\beta s_{lj} = \sum_{i=1}^{n} x_{iq} \sum_{A_j \ni \omega_l} |A_j|^{\alpha-1} m_{ij}^\beta \quad l = 1, c \quad q = 1, p, \tag{22}$$

and $\mathbf{H}$ a matrix of size $(c \times c)$ given by:

$$\mathbf{H}_{lk} = \sum_{i} \sum_{A_j \neq \emptyset} |A_j|^{\alpha-2} m_{ij}^\beta s_{lj} s_{kj} = \sum_{i} \sum_{A_j \supseteq \{\omega_k, \omega_l\}} |A_j|^{\alpha-2} m_{ij}^\beta \quad k, l = 1, c. \tag{23}$$

With these notations, V is solution of the following linear system:

$$\mathbf{H}V = \mathbf{B}, \tag{24}$$

which can be solved using a standard linear system solver. The way of deriving equations (20) to (24) from the optimality conditions of the problem is detailed in reference [20]. Note that, in practice, the resolution of system (24) is performed columnwise: each column of V is the solution of a linear system of c equations and c unknowns. As FCM and its variants, the algorithm starts with an initial guess for either the credal partition M or the cluster centers V and iterates until convergence, alternating the optimization of M and V.

2.4. Interpreting a credal partition

As underlined in [20], a credal partition is a rich representation that carries a lot of information about the data. In [20], various tools helping the user to interpret the results of ECM were suggested. First, a credal partition can be converted into classical clustering structures. For example, a fuzzy partition can be recovered by computing the pignistic probability $BetP_i(\{\omega_k\})$ induced by each bba m_i and interpreting this value as the degree of membership of object i to cluster k.

Another interesting way of synthesizing the information is to assign each object to the subset of classes with the highest mass. In this way, one obtains a partition in at most 2^c groups, which is referred to as a *hard credal partition*. This hard credal partition allows us to detect, on the one hand, the objects that can be assigned without ambiguity to a single cluster and, on the other hand, the objects lying at the boundary of two or more clusters.

It was also proposed to characterize each cluster by two sets of objects. The *lower approximation* ω_k^L of a cluster ω_k is the set of objects that belong with no doubt to cluster ω_k: it is the set of objects assigned to the singleton $\{\omega_k\}$ in the hard credal partition; the *upper approximation* ω_k^U gathers the objects that could *possibly* belong to cluster ω_k: it is the set of objects assigned to subsets of Ω containing ω_k.

Example 2. *Let us consider the credal partition presented in Table 1. The corresponding pignistic probabilities (using Yager's normalization) are given in Table 2. Lower and upper estimations of the clusters are:* $\omega_1^L = \{\mathbf{x}_1\}$, $\omega_2^L = \{\mathbf{x}_3\}$, $\omega_1^U = \{\mathbf{x}_1, \mathbf{x}_2\}$, $\omega_2^U = \{\mathbf{x}_3, \mathbf{x}_2\}$. *Object* $\mathbf{x}_4$ *is considered as an outlier.*

Table 2: Pignistic probabilities for the credal partition of Table 1

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
$BetP(\{\omega_1\})$	1	0.5	0.3	0.5
$BetP(\{\omega_2\})$	0	0.5	0.7	0.5

3. ECM with constraints

3.1. Expression of the constraints

Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be two objects associated with mass functions m_i and m_j. A mass function regarding the joint class membership of both objects may be computed from m_i and m_j in the Cartesian product $\Omega^2 = \Omega \times \Omega$. This mass function, denoted $m_{i\times j}$, is the combination of the vacuous extensions of m_i and m_j [25]. As shown in [10], it can be written as:

$$m_{i\times j}(A \times B) = m_i(A)\ m_j(B) \quad A, B \subseteq \Omega, A \neq \emptyset, B \neq \emptyset, \tag{25}$$

$$m_{i\times j}(\emptyset) = m_i(\emptyset) + m_j(\emptyset) - m_i(\emptyset)\ m_j(\emptyset). \tag{26}$$

From $m_{i\times j}$, we can compute the plausibility that the two objects $\mathbf{x}_i$ and $\mathbf{x}_j$ belong or not to the same class. In Ω^2, the event "Objects $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class" corresponds to the subset $\theta = \{(\omega_1, \omega_1), (\omega_2, \omega_2), \dots (\omega_c, \omega_c)\}$, whereas the event "Objects $\mathbf{x}_i$ and $\mathbf{x}_j$ do not belong to the same class" corresponds to its complement $\overline{\theta}$. The corresponding plausibilities are the following:

$$pl_{i\times j}(\theta) = \sum_{\{A\times B\subseteq\Omega^2 \mid (A\times B)\cap\theta\neq\emptyset\}} m_{i\times j}(A\times B) \quad (27)$$

$$= \sum_{A\cap B\neq\emptyset} m_i(A)\ m_j(B), \quad (28)$$

and

$$pl_{i\times j}(\overline{\theta}) = 1 - m_{i\times j}(\emptyset) - bel_{i\times j}(\theta), \quad (29)$$

$$= 1 - m_{i\times j}(\emptyset) - \sum_{\{A\times B\subseteq\Omega^2 \mid \emptyset\neq(A\times B)\subseteq\theta\}} m_{i\times j}(A\times B) \quad (30)$$

$$= 1 - m_{i\times j}(\emptyset) - \sum_{k=1}^{c} m_i(\{\omega_k\})\ m_j(\{\omega_k\}). \quad (31)$$

Example 3. *Let us consider a new collection of four objects to be classified into two classes. A credal partition, which expresses certain knowledge about the membership of the objects, is given in Table 3. Table 4 gives the mass functions of the joint membership of* $\mathbf{x}_1$ *with the three other objects. The associated plausibilities* $pl^{\Omega\times\Omega}(\theta)$ *and* $pl^{\Omega\times\Omega}(\overline{\theta})$ *are given in Table 5.*

Table 3: Credal partition to express constraints

A	$m_1(A)$	$m_2(A)$	$m_3(A)$	$m_4(A)$
$\emptyset$	0	0	0	0
$\{\omega_1\}$	1	1	0	0
$\{\omega_2\}$	0	0	1	0
Ω	0	0	0	1

This simple example shows how the joint membership of two objects may be represented using the plausibilities $pl^{\Omega\times\Omega}(\theta)$ and $pl^{\Omega\times\Omega}(\overline{\theta})$. In simple terms, the relevant information in Table 5 is contained in the zeros of these plausibilities. For example, nothing can be said about the joint membership of object $\mathbf{x}_1$ and $\mathbf{x}_4$, as both of these plausibilities are equal to 1. On the contrary, the fact that $pl^{\Omega\times\Omega}_{1\times 2}(\overline{\theta}) = 0$ indicates that ($\mathbf{x}_1$ and $\mathbf{x}_2$) are certainly in the same cluster. Equivalently, the null value of the plausibility $pl^{\Omega\times\Omega}_{1\times 3}(\theta)$ express the impossibility that $\mathbf{x}_1$ and $\mathbf{x}_3$ belong to the same class. These relationships will be used in the next section to propose a new formulation of ECM integrating pairwise constraints on instances.

Table 4: Masses of joint membership

$F = A \times B$	$m_{1\times 2}(F)$	$m_{1\times 3}(F)$	$m_{1\times 4}(F)$
$\{\omega_1\} \times \{\omega_1\}$	1	0	0
$\{\omega_1\} \times \{\omega_2\}$	0	1	0
$\{\omega_1\} \times \Omega$	0	0	1
$\{\omega_2\} \times \{\omega_1\}$	0	0	0
$\{\omega_2\} \times \{\omega_2\}$	0	0	0
$\{\omega_2\} \times \Omega$	0	0	0
$\Omega \times \{\omega_1\}$	0	0	0
$\Omega \times \{\omega_2\}$	0	0	0
$\Omega \times \Omega$	0	0	0

Table 5: Plausibilities for the events θ and $\overline{\theta}$

F	$pl_{1\times 2}(F)$	$pl_{1\times 3}(F)$	$pl_{1\times 4}(F)$
θ	1	0	1
$\overline{\theta}$	0	1	1

3.2. Objective function of CECM

Let us now assume that the credal partition is unknown and that we are given some pairwise constraints. As explained in the introduction, we consider that these constraints are must-link or cannot-link constraints. Let $\mathcal{M}$ denote the set of pairs of objects constrained by a must-link and $\mathcal{C}$ the set of pairs of objects constrained by a cannot-link. One has to seek for a credal partition that reflects both the similarities computed from the data and the constraints. A natural requirement is that $pl_{i\times j}(\theta)$ be as low as possible if $(\mathbf{x}_i, \mathbf{x}_i) \in \mathcal{C}$. In the same way, $(\mathbf{x}_i, \mathbf{x}_i) \in \mathcal{M}$ implies that $pl_{i\times j}(\overline{\theta})$ should be as low as possible. To achieve this goal, we suggest to integrate penalty terms into the ECM criterion and we propose to minimize the following objective function:

$$J_{\text{CECM}}(M, V) = J_{\text{ECM}}(M, V) + \frac{\gamma}{|\mathcal{M}|} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M}} pl_{i\times j}(\overline{\theta}) + \frac{\eta}{|\mathcal{C}|} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C}} pl_{i\times j}(\theta), \tag{32}$$

such that the constraints (19) are respected. The second and third terms represent, respectively, the cost of violating the must-link the cannot-link constraints. The coefficients γ and η control the tradeoff between the objective function of ECM and the constraints.

3.3. Optimization

As in FCM, NC and ECM, we propose an alternate optimization scheme in order to fix the partition matrix M and the centroid matrix V. First, we

note that the two penalty terms added to the objective function of ECM do not depend on the cluster centroids. The same update scheme for the centroids (equations (22) to (24)) can thus be used in CECM.

Generally, the problem is much more complex for the belief masses, and a direct update equation of the m_{ij} from the optimality conditions like (20) is no longer possible. However, if we fix $\beta = 2$ then the objective function (32) becomes:

$$J_{\text{CECM}}(M,V) = \sum_{i=1}^{n}\sum_{A_k\neq\emptyset}|A_k|^{\alpha}m_{ik}^2d_{ik}^2 + \sum_{i=1}^{n}\rho^2m_{i\emptyset}^2 + \frac{\gamma}{|\mathcal{M}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}}pl_{i\times j}(\overline{\theta}) + \frac{\eta}{|\mathcal{C}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}}pl_{i\times j}(\theta). \quad (33)$$

Because of Equation (29), Equation (33) is equal to:

$$J_{\text{CECM}}(M,V) = \sum_{i=1}^{n}\sum_{A_k\neq\emptyset}|A_k|^{\alpha}m_{ik}^2d_{ik}^2 + \sum_{i=1}^{n}\rho^2m_{i\emptyset}^2 - \frac{\gamma}{|\mathcal{M}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}}m_{i\times j}(\emptyset) - \frac{\gamma}{|\mathcal{M}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}}bel_{i\times j}(\theta) + \frac{\eta}{|\mathcal{C}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}}pl_{i\times j}(\theta) + \gamma. \quad (34)$$

Note that the last term of equation (34), which is constant, will be omitted in the rest of the paper. It can be seen that the objective function is, in that case, quadratic with respect to the m_{ij}. To make this point clearer, let $\mathbf{m}_i$ denote the vector of size 2^c of the masses related to object $\mathbf{x}_i$. Let $\Phi^i = (\phi^i_{kl})$ be n diagonal matrices ($i = 1, n$) of size ($2^c \times 2^c$) defined by:

$$\phi^i_{kl} = \begin{cases} \rho^2 & \text{if } A_k = A_l = \emptyset, \\ d_{ik}^2|A_k|^{\alpha} & \text{if } k = l \text{ and } A_k \neq \emptyset, \\ 0 & \text{else.} \end{cases} \quad (35)$$

Let us also define two matrices $\Delta^{\mathcal{M}} = (\delta^{\mathcal{M}}_{kl})$ and $\Delta^{\mathcal{C}} = (\delta^{\mathcal{M}}_{kl})$ of size ($2^c \times 2^c$) as follows:

$$\delta^{\mathcal{M}}_{kl} = \begin{cases} 1 & \text{if } A_k = \emptyset \text{ or } A_l = \emptyset, \\ 1 & \text{if } A_k = A_l \text{ and } |A_k| = |A_l| = 1, \\ 0 & \text{else.} \end{cases} \quad (36)$$

$$\delta^{\mathcal{C}}_{kl} = \begin{cases} 1 & \text{if } A_k \cap A_l \neq \emptyset \\ 0 & \text{else.} \end{cases} \quad (37)$$

With these notations, J_{CECM} may be rewritten as follows:

$$J_{\text{CECM}}(M,V) = \sum_{i=1}^{n}\mathbf{m}_i^t\Phi^i\mathbf{m}_i - \frac{\gamma}{|\mathcal{M}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}}\mathbf{m}_i^t\Delta^{\mathcal{M}}\mathbf{m}_j + \frac{\eta}{|\mathcal{C}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}}\mathbf{m}_i^t\Delta^{\mathcal{C}}\mathbf{m}_j, \quad (38)$$

As we have linear constraints, a classical optimization method [30] can be used and the convergence is insured in a reasonable time. The overall procedure is summarized in Algorithm 1.

Algorithm 1 CECM with an Euclidean metric

Input: Number c of desired clusters, n objects $\mathbf{x}_1, ..., \mathbf{x}_n$, set of cannot-link $\mathcal{C}$, set of must-link $\mathcal{M}$
Output: Credal partition matrix M, centroid matrix V
Random Initialization of V
repeat
1) Calculate the new masses by solving the quadratic programming problem defined by (38) subject to (19).
2) Calculate the new centroids by solving the linear system defined by equations (22) to (24).
until No significant change in V between two successive iterations

4. CECM with an adaptive metric

4.1. Model

In the ECM algorithm, the distance d_{ik}^2 between the object $\mathbf{x}_i$ and the centroid $\overline{\mathbf{v}}_k$ is a Euclidean distance. Classes are then supposed to be spherical. However, the use of a Mahalanobis distance may be interesting in case of elliptical clusters. Using an adaptive metric can be highly desirable when using constraints, in particular when these constraints contradict a Euclidean model. To modify the previous algorithm, we follow an approach inspired from Gustafson and Kessel [13] and well described in [14]. Let S_l denote a $(p \times p)$ matrix associated to cluster ω_l $(l = 1, c)$ inducing a norm $||\mathbf{x}||_{S_l}^2 = \mathbf{x}^t S_l \mathbf{x}$. Using the same approach that we used for the centroids, we compute the matrix $\overline{S}_j$ associated with a non singleton A_j by averaging the matrices associated to the classes $\omega_k \in A_j$:

$$\overline{S}_j = \frac{1}{|A_j|} \sum_{l=1}^{c} s_{lj} S_l, \quad \forall A_j \subseteq \Omega, A_j \neq \emptyset. \tag{39}$$

The distance d_{ij}^2 between $\mathbf{x}_i$ and any set $A_j \neq \emptyset$ is then defined by:

$$d_{ij}^2 = ||\mathbf{x}_i - \overline{\mathbf{v}}_j||_{\overline{S}_j}^2 = (\mathbf{x}_i - \overline{\mathbf{v}}_j)^t \overline{S}_j (\mathbf{x}_i - \overline{\mathbf{v}}_j). \tag{40}$$

The new criterion to be minimized thus becomes:

$$J_{\text{CECM}}(M, V, S_1, \cdots, S_c) = \sum_{i=1}^{n} \sum_{A_k \neq \emptyset} |A_k|^{\alpha} m_{ik}^2 ||\mathbf{x}_i - \overline{\mathbf{v}}_j||_{\overline{S}_j}^2 + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^2 + \frac{\gamma}{|\mathcal{M}|} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M}} pl_{i \times j}(\overline{\theta}) + \frac{\eta}{|\mathcal{C}|} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C}} pl_{i \times j}(\theta). \tag{41}$$

4.2. Optimization

We first note that the minimization of (41) with respect to the masses is independent of the metric, so that the way of deriving the masses by a constrained quadratic optimization is unchanged. In their algorithm, Gustafson

and Kessel showed that the update equations of FCM for the cluster centers were not affected by the introduction of a metric associated to each cluster. On the contrary, in CECM, the determination of the centers takes explicitly into account the metric, as shown below.

4.2.1. Optimization with respect to the cluster centers

We first consider that M and the matrices S_l $(l = 1, c)$ are fixed. The minimization of J_{CECM} with respect to V is an unconstrained optimization problem. The partial derivatives of J_{CECM} with respect to the centers are given by:

$$\frac{\partial J_{\text{CECM}}}{\partial \mathbf{v}_l} = \sum_{i=1}^{n} \sum_{A_j \neq \emptyset} |A_j|^{\alpha} m_{ij}^2 \frac{\partial d_{ij}^2}{\partial \mathbf{v}_l} \quad l = 1, c. \tag{42}$$

$$\frac{\partial d_{ij}^2}{\partial \mathbf{v}_l} = 2(s_{lj})\overline{S}_j(\mathbf{x}_i - \bar{\mathbf{v}}_j)(-\frac{1}{|A_j|}) \quad l = 1, c. \tag{43}$$

From (42) and (43) we thus have:

$$\frac{\partial J_{\text{CECM}}}{\partial \mathbf{v}_l} = -2\sum_{i=1}^{n} \sum_{A_j \neq \emptyset} |A_j|^{\alpha-1} m_{ij}^2 s_{lj} \overline{S}_j(\mathbf{x}_i - \bar{\mathbf{v}}_j) \tag{44}$$

$$= -2\sum_{i=1}^{n} \sum_{A_j \neq \emptyset} |A_j|^{\alpha-1} m_{ij}^2 s_{lj} \overline{S}_j(\mathbf{x}_i - \frac{1}{|A_j|}\sum_{k} s_{kj}\mathbf{v}_k) \quad l = 1, c. \tag{45}$$

Setting these derivatives to zero gives l equations in $\mathbf{v}_k$ which can be written as:

$$\sum_{i} \sum_{A_j \neq \emptyset} |A_j|^{\alpha-1} m_{ij}^2 s_{lj} \overline{S}_j \mathbf{x}_i = \sum_{k} \sum_{i} \sum_{A_j \neq \emptyset} |A_j|^{\alpha-2} m_{ij}^2 s_{lj} s_{kj} \overline{S}_j \mathbf{v}_k \quad l = 1, c, \tag{46}$$

or, equivalently:

$$\sum_{i} \sum_{A_j \ni \omega_l} |A_j|^{\alpha-1} m_{ij}^2 \overline{S}_j \mathbf{x}_i = \sum_{k} \sum_{i} \sum_{A_j \supseteq \{\omega_k, \omega_l\}} |A_j|^{\alpha-2} m_{ij}^2 \overline{S}_j \mathbf{v}_k \quad l = 1, c. \tag{47}$$

Let $\mathbf{F}^{(l,i)}$ denote the $(p \times p)$ matrix:

$$\mathbf{F}^{(l,i)} = \sum_{A_j \ni \omega_l} |A_j|^{\alpha-1} m_{ij}^2 \overline{S}_j \quad l = 1, c \quad i = 1, n, \tag{48}$$

and $\mathbf{G}^{(l,k)}$ denote the $(p \times p)$ matrix:

$$\mathbf{G}^{(l,k)} = \sum_{i} \sum_{A_j \supseteq \{\omega_k, \omega_l\}} |A_j|^{\alpha-2} m_{ij}^2 \overline{S}_j \quad k, l = 1, c. \tag{49}$$

Next, we form, from these two $(p \times p)$ matrices, two new matrices $\mathbf{F}$ and $\mathbf{G}$, of size $(cp \times np)$ and $(cp \times cp)$, respectively:

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}^{(1,1)} & \mathbf{F}^{(1,2)} & \dots & \mathbf{F}^{(1,n)} \\ \mathbf{F}^{(2,1)} & \mathbf{F}^{(2,2)} & \dots & \mathbf{F}^{(2,n)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{F}^{(c,1)} & \mathbf{F}^{(c,2)} & \dots & \mathbf{F}^{(c,n)} \end{pmatrix} \quad \mathbf{G} = \begin{pmatrix} \mathbf{G}^{(1,1)} & \mathbf{G}^{(1,2)} & \dots & \mathbf{G}^{(1,c)} \\ \mathbf{G}^{(2,1)} & \mathbf{G}^{(2,2)} & \dots & \mathbf{G}^{(2,c)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}^{(c,1)} & \mathbf{G}^{(c,2)} & \dots & \mathbf{G}^{(c,c)} \end{pmatrix} \tag{50}$$

Let us stack all object $\mathbf{x}_i$ in a same vector $\mathbf{X}$ of size $(np \times 1)$ and rearrange matrix V in the form of a vector of size $(cp \times 1)$ such that:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \quad V = \begin{pmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_c \end{pmatrix}$$

With all these notations, vector V is solution of the following linear system:

$$\mathbf{G}V = \mathbf{F}\mathbf{X}. \tag{51}$$

We see that, instead of solving p system of c unknowns as in the case of a Euclidean metric, we have to solve a unique system of cp equations and cp unknowns. This higher complexity is the price to pay for an automatic adaptation of the metric.

4.2.2. Optimization with respect to the metrics S_l

We now consider that M and V are fixed and we want to determine the matrices S_l. We follow the same line of reasoning as Gustafson and Kessel. In order to avoid the degenerate solution consisting of matrices S_l with zero entries, we impose that the clusters have a constant volume using the constraints $\det(S_l) = 1$ for all $l = 1, c$. To solve the constrained minimization problem with respect to $S_1, \cdots, S_c$, we introduce c Lagrange multipliers λ_i and write the Lagrangian:

$$\mathcal{L}(S_1, \cdots, S_c, \lambda_1, \cdots, \lambda_c) = J_{\text{CECM}}(M, V) - \sum_{k=1}^{c} \lambda_k \left(\det(S_k) - 1\right) \tag{52}$$

We recall that the definition of the distance of an object $\mathbf{x}_i$ to a focal set A_j is:

$$d_{ij}^2 = (\mathbf{x}_i - \overline{\mathbf{v}}_j)^t \overline{S}_j (\mathbf{x}_i - \overline{\mathbf{v}}_j) = (\mathbf{x}_i - \overline{\mathbf{v}}_j)^t \left(\frac{1}{|A_j|} \sum_{k=1}^{c} s_{kj} S_k \right) (\mathbf{x}_i - \overline{\mathbf{v}}_j). \tag{53}$$

Starting from the fact that the derivatives of $\mathbf{x}^t A \mathbf{x}$ and $\det(A)$ with respect to a symmetric matrix A are $\mathbf{x}\mathbf{x}^t$ and $\det(A)A^{-1}$ respectively, we obtain the following derivative of $\mathcal{L}$ with respect to matrix S_l:

$$\frac{\partial \mathcal{L}}{\partial S_l} = \sum_i \sum_{A_j \neq \emptyset} m_{ij}^2 |A_j|^{\alpha-1} s_{lj} (\mathbf{x}_i - \overline{\mathbf{v}}_j)(\mathbf{x}_i - \overline{\mathbf{v}}_j)^t - \lambda_l \det(S_l) S_l^{-1} \quad l = 1, c. \tag{54}$$

The derivatives with respect to the Lagrange multipliers lead to the constraints $\det(S_l) = 1$ for all i. Let Σ_l denote the following matrix:

$$\Sigma_l = \sum_i \sum_{A_j \ni \omega_l} m_{ij}^2 |A_j|^{\alpha-1} (\mathbf{x}_i - \overline{\mathbf{v}}_j)(\mathbf{x}_i - \overline{\mathbf{v}}_j)^t \quad l = 1, c. \tag{55}$$

Note that Σ_l can be considered as the analog in the evidential framework of the fuzzy covariance matrix. From (54), we have:

$$\Sigma_l = \lambda_l S_l^{-1} \quad l = 1, c, \tag{56}$$

and, thus

$$\Sigma_l S_l = \lambda_l I \quad l = 1, c, \tag{57}$$

where I denote the $(p \times p)$ identity matrix. Taking the determinant of this last equation leads to:

$$\det(\Sigma_l S_l) = \det(\Sigma_l) \det(S_l) = \det(\Sigma_l) = \lambda_l^p \quad l = 1, c. \tag{58}$$

It follows that

$$\lambda_l = \det(\Sigma_l)^{\frac{1}{p}} \quad l = 1, c. \tag{59}$$

Replacing λ_l by its expression and using (56), we finally obtain:

$$S_l = \det(\Sigma_l)^{\frac{1}{p}} \Sigma_l^{-1} \quad l = 1, c. \tag{60}$$

Note that Σ_l is invertible since it is symmetric and positive definite. Indeed, each $(\mathbf{x}_i - \overline{\mathbf{v}}_j)(\mathbf{x}_i - \overline{\mathbf{v}}_j)^t$ is symmetric, positive and semi-definite, and so is their weighted sum.

The overall CECM procedure with an adaptive metric is summarized in Algorithm 2.

Algorithm 2 CECM with an adaptive metric

Input: Number c of desired clusters, n objects $\mathbf{x}_1, ..., \mathbf{x}_n$, set of cannot-link $\mathcal{C}$, set of must-link $\mathcal{M}$

Output: Credal partition matrix M, centroid matrix V, set of matrices S_l $l = 1, c$

Random Initialization of V

repeat

1) Calculate the new masses by solving the quadratic programming problem defined by (38) subject to (19).

2) Calculate the new centroids by solving the linear system defined by equations (48) to (51).

3) Calculate the new matrices S_l, $l = 1, c$ using (55) and (60).

until No significant change in V between two successive iterations

5. Experimental results

5.1. Datasets

The performances of CECM were evaluated on three data sets. In order to illustrate the interest of introducing constraints, we created a synthetic dataset, represented in Figure 1. It consists in two classes of patterns in a two-dimensional space. In each class, patterns were generated according to a mixture of two Gaussians, with means $(0, 0)$ and $(0, 7)$ in the first class, and $(7, 0)$ and $(7, 7)$ in the second one. All the Gaussians have a common covariance matrix $2\,\mathrm{Id}_2$, where Id_2 denotes the identity matrix in R^2. In the two classes, the proportions of the Gaussians are the same: 100 points were drawn from each.

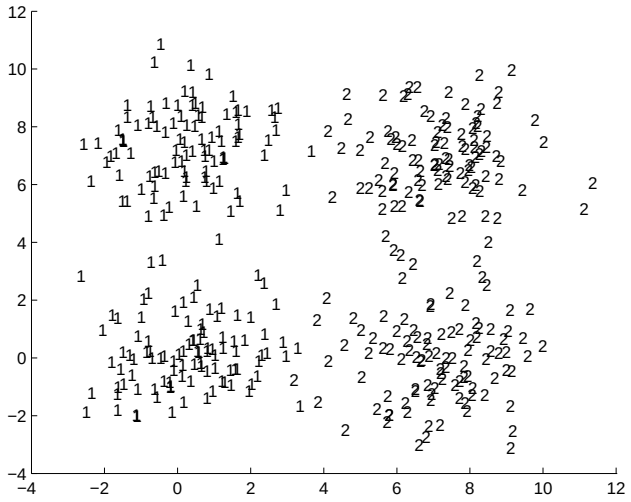


Figure 1: Synthetic data set

The Iris data set is composed of three classes in a four-dimensional space. Each one contains fifty samples. We stress here that only the setosa class is linearly separable from the others. Moreover, classes have non-spherical distributions, which suggests that the Mahalanobis distance may be best suited than the Euclidean one to process these data. In the Ionosphere data set, 351 patterns are separated into two classes of 225 patterns and 126 patterns, respectively. Each pattern is described by 34 attributes. Both the Iris and Ionosphere data sets may be downloaded from the UCI Machine Learning Repository (http://www.ics.uci.edu/~mlearn).

5.2. Comparing two partitions

In order to evaluate the accuracy of a clustering algorithm, the crisp partition $\widehat{P}$ found may be compared to some reference partition P. Remark that this task is not trivial. Since the label of each cluster is arbitrary and does not reflect any ground truth (unlike in supervised classification), two identical partitions may label differently same groups of data.

To overcome this difficulty, various methods have been proposed to compute a degree of similarity between P and $\widehat{P}$. Let a (respectively, b) be the number of pairs of objects simultaneously assigned to identical classes (respectively, different classes) in P and $\widehat{P}$. The Rand Index (RI) estimates the degree of global compatibility between P and $\hat{P}$ by:

$$RI(P, \widehat{P}) = \frac{2(a+b)}{n(n-1)}. \tag{61}$$

Remark that with CECM, $\widehat{P}$ was determined by assigning each object to the cluster with maximal pignistic probability after convergence of the algorithm.

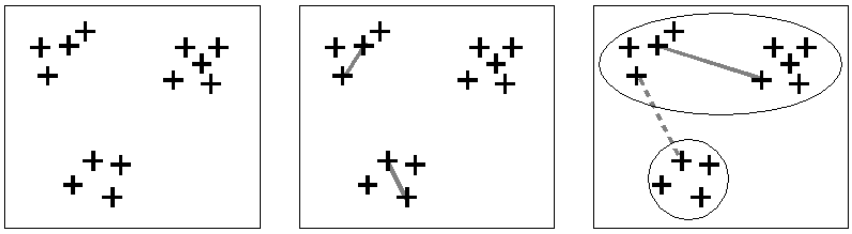
5.3. Choice of the parameters

We address here the choice of the parameters used to run the various experiments. First of all, the number c of classes was defined by the user. In order to obtain a significant level of non-specificity, so that the credal partition computed differs from a fuzzy partition, parameter α was set to 1. The values of parameter ρ differ according to the data processed; therefore, they will be indicated throughout the presentation of the results. To give the same importance to must-link and cannot-link constraints, we set $\gamma = \eta$.

5.4. Choice of the constraints

Constraints were defined using two different methods. Random selection consists in randomly selecting two patterns in the dataset. Then, the true relationship between these points is identified using the true partition of the data. This technique allows us to introduce a high number of constraints, and thus to study the behaviour of the algorithm in various situations.

However, in some applications, the constraints may not always be available a priori, but an oracle (a user) may be available to provide the constraints. This scheme, where the system queries the oracle to obtain information is called *active learning* [1, 6]. It is trivial to notice that, among the possible pairwise constraints, some of them are informative with respect to the clustering problem, while some of them are useless, as illustrated in Figure 2. Additionally, several authors have observed that a constrained clustering approach with a bad choice of pairs can deteriorate the clustering performances [9, 26]. The goal of an active learning is thus to select the pairwise constraints which are the most informative about the underlying structure of the objects, so that the clustering performance can be improved with as few queries as possible. We propose to introduce constraints incrementally by alternatively running CECM, selecting



(a) (b) (c)

Figure 2: Pairwise constraints patterns for a dataset (a): some are useless (b) and some are informative (c) to lead the algorithm towards a desired solution.

pairs of objects, and asking an expert to identify the nature of the corresponding constraints, until a specified number of constraints is reached. We use the credal partition obtained with CECM to select the most suitable pairs of objects. These pairs are selected according to the following requirements:

- The first object must be classified with a high degree of uncertainty,
- The second object must be classified with a high degree of certainty.

Indeed, if the uncertainty about the membership of the two objects is low, the constraint may be non informative and conversely, if the uncertainty regarding the classification of both objects is high, the constraint may lead to misclassify both objects. Different ways to find such objects thanks to the credal partition or to the centroids can be considered. We propose a strategy that proved experimentally to be efficient. The points for which the uncertainty is high are the points assigned in the hard credal partition (see Section 2.4) to focal sets of cardinality greater than 1. In particular, points associated to focal sets A_j such that $A_j = 2$ are likely to be located at the boundary of two clusters. Thus, for the selection of the first object, we propose to select the point associated to the highest mass allocated to focal sets of cardinality equal to 2. For the second object, we pick up the nearest point from one of the centroids. The user is then provided with this pair of points, and enters either a must-link or a cannot-link constraint.

5.5. Results on synthetic and real data

5.5.1. Interest of adding constraints

We first illustrate the interest of introducing constraints using the synthetic data set. The ECM algorithm was run using a Euclidean distance, with $\rho^2 = 100$. The credal partition obtained shows a diagonal boundary between the two classes. The direction of the boundary (from upper left to lower right, or from lower left to upper right) depends on the initialization of the centroids. Figure 3 represents one of the credal partitions obtained. Here, each point is associated with the non-empty subset $A \subseteq \Omega$ that received the highest amount of belief

mass. The two large crosses represent the centroids obtained after convergence. The RI is equal to 0.56.

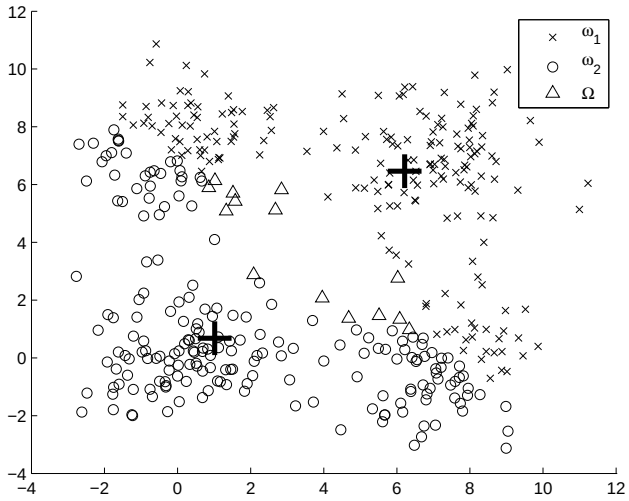


Figure 3: Hard credal partition obtained using ECM with a Euclidean metric.

The Euclidean distance implicitly supposes that the classes are spherical, which is obviously not the case for this data set. If we use the Mahalanobis distance instead (with the same parameter values as before), we obtain either an horizontal or a vertical boundary between the classes. Figure 4 shows one of the solutions obtained, where the boundary is horizontal. In this case, the credal partition does not correspond to the true partition of the data, and the RI is equal to 0.5.

The add of a small number of randomly chosen constraints allows us to lead the algorithm towards the desired solution. For example, by using only ten constraints, CECM finds the desired classes, as it is shown in Figure 5. Here, a solid line segment between two points corresponds to a must-link constraint between two objects and a dashed line segment between two points corresponds to a cannot-link constraint.

5.5.2. Influence of the penalty coefficients γ and η

As pointed out in Section 5.3, choosing adequate values for parameters γ and η may be difficult. If the values are too high, satisfying the constraints prevails over finding compact classes. Conversely, too low values may lead to ignoring the constraints. Figure 6 shows the average RI, obtained on the Iris data set using an adaptive metric, plotted against the number of constraints. For each number of constraints, this average was computed over 100 different

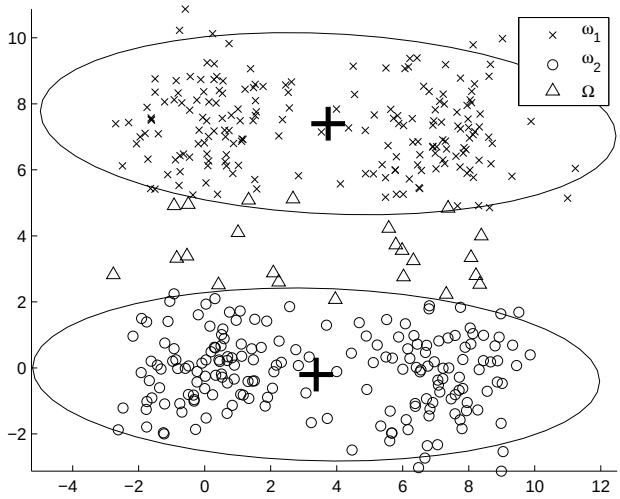


Figure 4: Hard credal partition obtained using ECM with an adaptive metric.

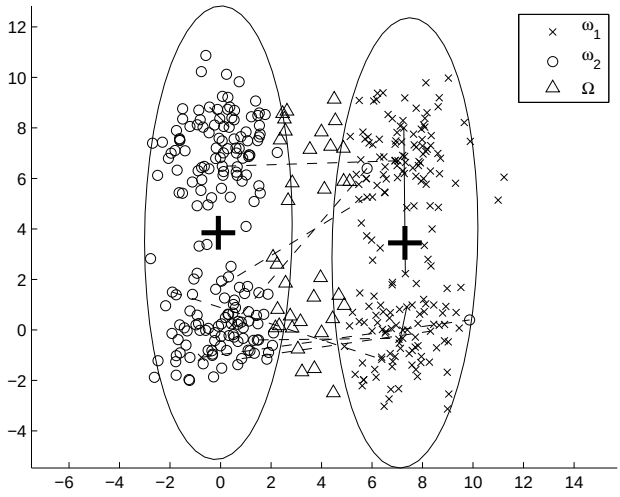


Figure 5: Hard credal partition obtained using CECM with an adaptive metric and 10 constraints; solid lines represent must-link constraints, and dashed lines cannot-link constraints.

classifications obtained with CECM using an adaptive metric. Remark that the respective rates of must-link and cannot-link constraints are randomly chosen for each run of the algorithm. Unsurprisingly, it may be noticed that the accuracy

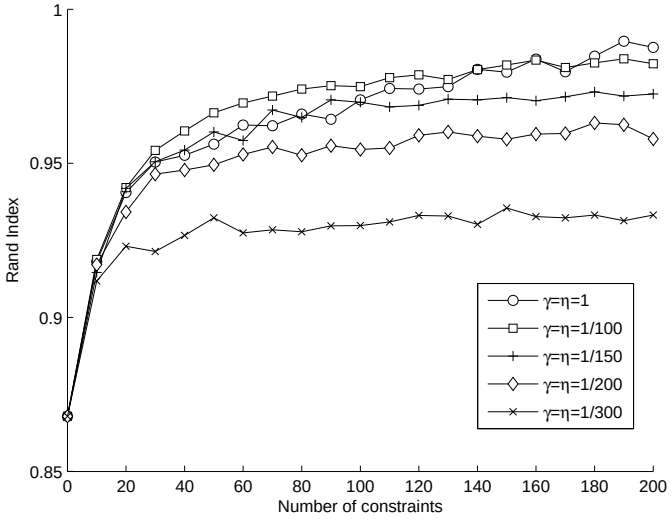


Figure 6: Average Rand Index as a function of the number of constraints, Iris data set.

of the classification increases with the number of constraints introduced. Note that the results obtained on the Iris data set are similar to those presented in [12].

Throughout the experiments, we remarked that high penalty coefficients may sometimes yield worse results than low coefficients when the number of constraints is low, but better results when the number of constraints increases. For example, it may be noticed in Figure 6 that the results obtained with $\gamma = \eta = 1/100$ are better or equal to those obtained with $\gamma = \eta = 1$, when up to 170 pairwise constraints are taken into account. For 180, 190 or 200 constraints, the parameter values $\gamma = \eta = 1$ give the best results. This behaviour may be explained as follows. It is likely that a small set of constraints covers scarce regions in the input space. Thus, enforcing these constraints introduce inconsistencies in the partition. Indeed, the mass functions associated with the constrained points may be modified, so that the class centers move in undesired directions or on inadequate distances. If the number of constraints increases, the coverage of the input space will likely increase as well. If instead few constraints are available and the values of γ and η are low, finding compact classes may prevail over respecting the constraints, yielding again a credal partition with good consistency. In particular, this explains why highly penalizing the violation of constraints decreases the accuracy of the partition when few con-

straints are available. This well-known behaviour has often been observed on constrained algorithms derived from the hard c-means algorithm [26, 8].

5.5.3. Random selection of constraints and active learning

Here, we examine the behaviour of the algorithm and make a comparison between an active learning scheme and a random constraint selection method. The experiments were conducted on the Iris and Ionosphere data sets. We used parameter values $\gamma = \eta = 1$ and $\rho^2 = 1000$ for both data sets.

Figures 7 and 8 show the evolution of the average RI (computed over 100 trials) according to the number of pairwise constraints, for both data sets. The pairwise constraints were randomly selected. The average RI is computed both over all objects and over unconstrained objects.

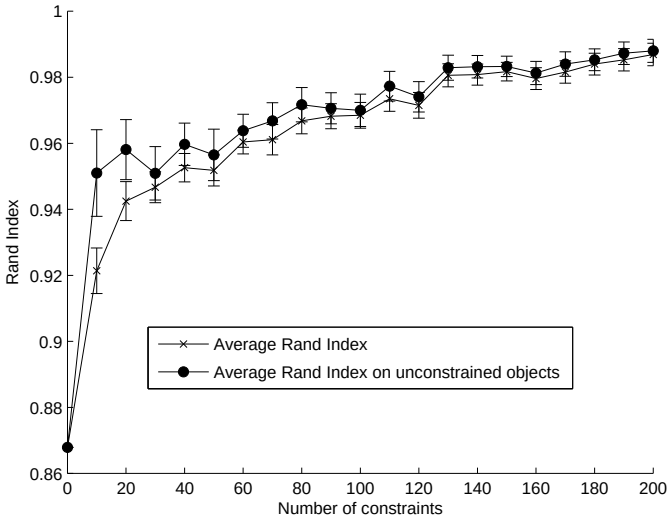


Figure 7: Average Rand Index as a function of the number of randomly selected constraints (Iris data set).

We remark that the RI computed over unconstrained objects increases with the number of constraints. Therefore, introducing constraints allows us to guide the algorithm towards a better solution, and does not only improve the classification of constrained objects. Remark that the RI computed over constrained objects may increase with the number of constraints. The reason is that a constraint involving a data point misclassified with a high degree of belief may have a negative effect on the classification. Indeed, in this case, the centers of the classes may move in undesired directions, and the other constrained point, previously well classified, may switch to the wrong class.

Active learning, being a way of introducing constraints on carefully selected points, seems a good way of avoiding such a situation. Figures 9 and 10 compare

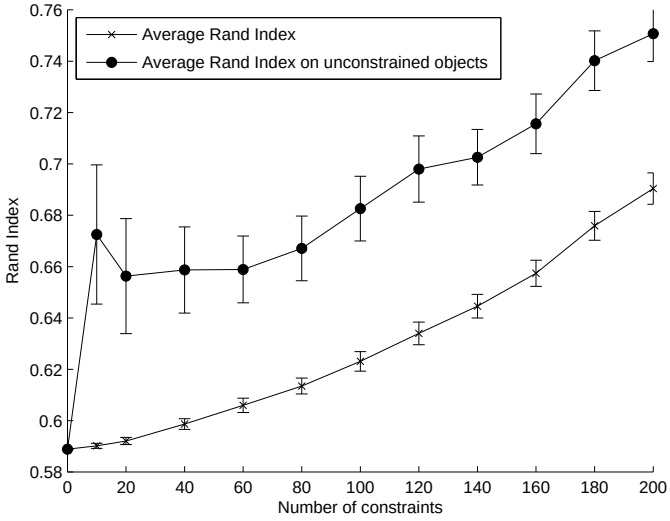


Figure 8: Average Rand Index as a function of the number of randomly selected constraints (Ionosphere data set).

the active learning scheme with random selection of constraints. Overall, active learning allows faster convergence than does random selection of constraints. In the case of the Iris dataset, the optimum is obtained with 40 constraints when using active learning, whereas it is still not obtained with 200 constraints when using random selection. Remark that active learning may be outperformed by random selection, especially with a low number of constraints. In this case, active learning tends to introduce constraints on data that belong to specific regions of the input space. As explained in Section 5.5.2, this may result in undesired moves of the class centers. As a consequence, other data points whose distances to the center increase may switch to other classes.

5.6. Application to medical image segmentation

The interest of CECM will now be illustrated using an example in medical imaging taken from [5]. An image of a pathological brain was acquired using magnetic resonance imaging. It is represented in Figure 11. In this image, according to the gray levels of the pixels, three main areas may be distinguished: the brightest area corresponds to the pathological area, the dark gray to normal brain tissues and intermediate gray levels correspond to ventricles and cerebrospinal fluid. The aim was to isolate the tumor from the other parts of the brain by looking for a partition into c=2 clusters. To make the computations tractable, the gray levels of the 156x141 pixels of the original image were quantified into 400 prototypes using a basic learning vector quantization algorithm

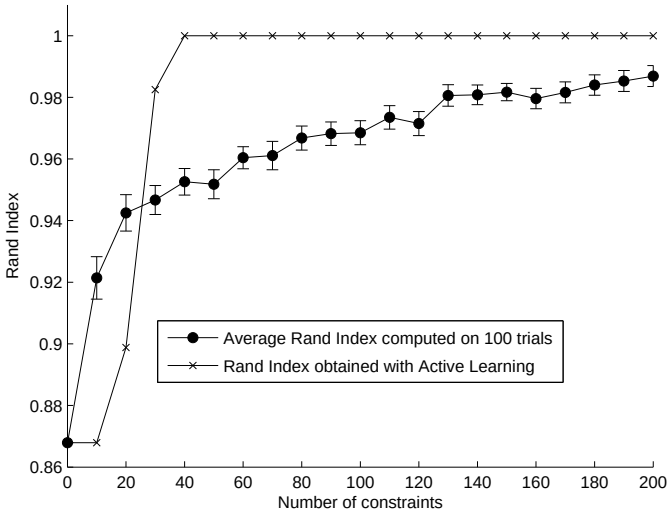


Figure 9: Rand Index obtained using Active learning, and average Rand Index obtained using randomly selected constraints, as a function of the number of constraints (Iris data set).

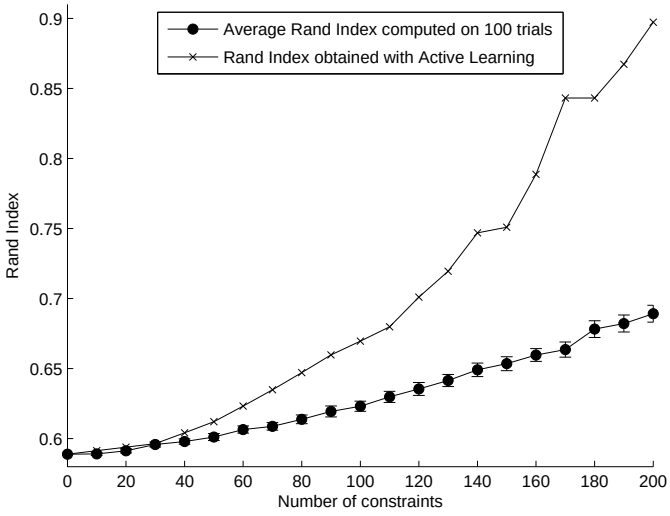


Figure 10: Rand Index obtained using Active learning, and average Rand Index obtained using randomly selected constraints, as a function of the number of constraints (Ionosphere data set).

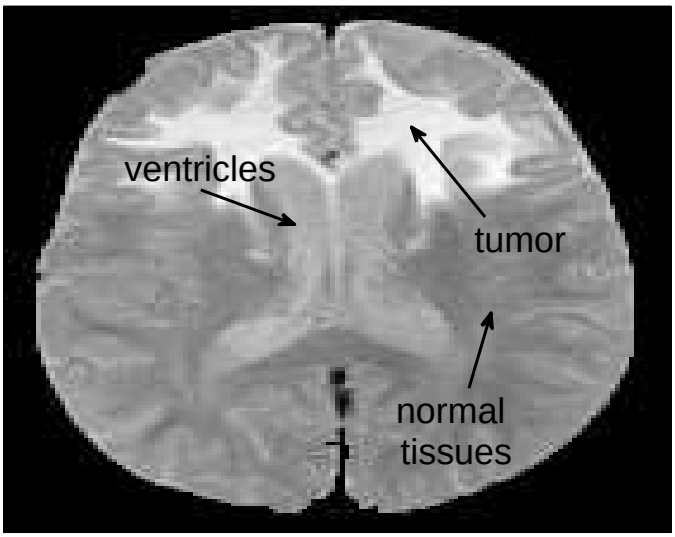


Figure 11: Original image of the brain; the bright area corresponds to a tumor; the dark gray one, to normal brain tissues, and the intermediate gray one to ventricles and cerebrospinal fluid.

[16]. The clustering was performed on this set of prototypes and the pixels in the image were assigned to the class of the nearest prototype.

Starting from the gray levels of the pixels (rescaled between 0 and 1), ECM, with $c = 2$, $\alpha = 2$, and $\rho^2 = 10$, finds the hard credal partition represented in Figure 12. White and light grays represent two clusters and the darker gray is given to pixels assigned to Ω in the hard credal partition. In a next experiment, imitating what could be done by an expert, we introduced constraints as indicated in Figure 13. White areas corresponds to pixels related by a must-link and these two areas are mutually linked by a cannot-link. The hard credal partition obtained by applying CECM with the adaptive metric (with $\gamma = \eta = 0.01$ and $\alpha = 2$, $\rho 2 = 10$) is shown in Figure 14. It may be seen that the constraints made it possible to raise the indetermination concerning the pixels allocated to Ω and thus to properly isolate the pathological area. As a matter of comparison, the partitions computed from the pignistic probabilities obtained by ECM and CECM are given in Figure 15.

6. Conclusion

In this paper, we addressed the problem of introducing constraints in a classification task. Our work is based within the theoretical framework of belief functions. In this framework, the ECM algorithm computes a credal partition of the data: each pattern is associated with a belief function that describes its

Figure 12: Hard credal partition obtained from ECM with an Euclidean metric (white: ω_1, light gray: ω_2, dark gray: Ω).

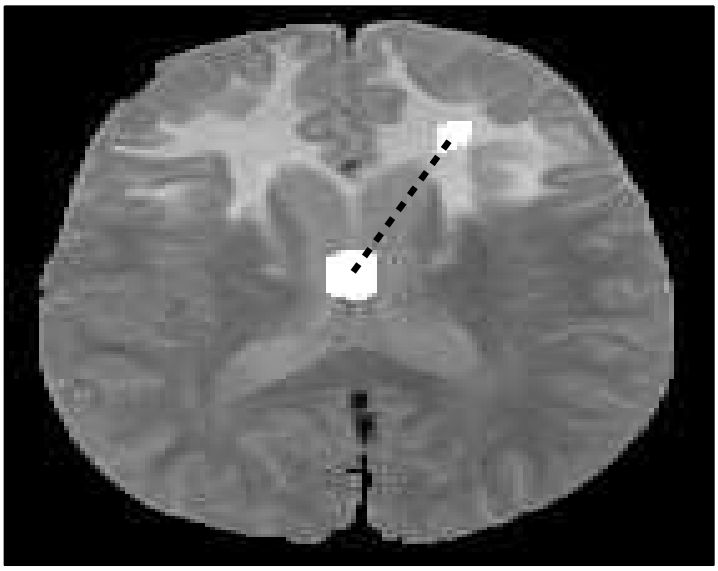


Figure 13: Must-link constraints (white areas) and cannot-link constraint (dashed line) introduced by an expert.

Figure 14: Hard credal partition obtained from CECM with an adaptive metric (white: ω_1, light gray: ω_2, dark gray: Ω).

membership to the classes. Our contribution is twofold. We introduced the Mahalanobis distance in the ECM algorithm, in order to handle non-spherical classes. We also presented an extension of the ECM algorithm, called CECM, which takes additional information into account in the clustering process. This information takes the form of pairwise constraints between the data: a must-link constraint indicates that two patterns must be classified into the same class; a cannot-link constraint, that they must be classified into different classes. We proposed an active-learning procedure, in which an expert is questioned about the relationships between pairs of data. Selecting these pairs is obviously a crucial issue for introducing relevant constraints. In our algorithm, the selection step may be easily conducted using the semantics of belief functions.

Our experiments show that introducing constraints improves the accuracy of the partition obtained, by guiding the algorithm towards desired solutions. When complex models are used, such as the Mahalanobis metric for computing distances between data, constraints allow us to compute parameter estimates that better fit the problem considered. We also showed that the number of constraints required to obtain an accurate clustering of the data need not be huge. In particular, much fewer constraints were necessary to reach the optimal partition when using our active-learning procedure than when constraints were randomly chosen. We also studied the influence of the constraints on the consistency of the solution obtained. Finally, we demonstrated the interest of our approach by on a medical image segmentation problem. The aim was to

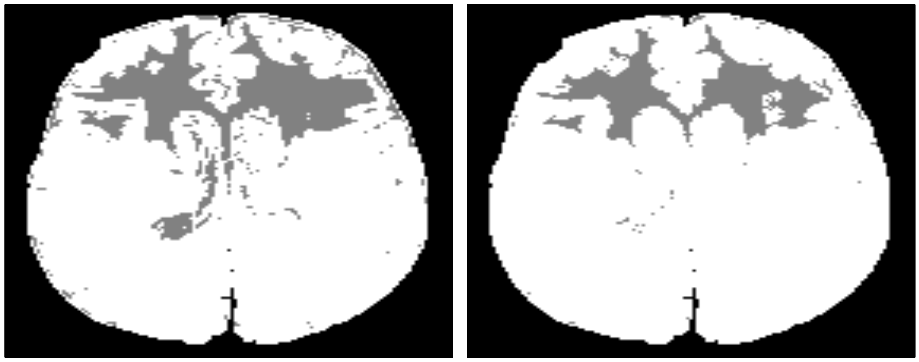


Figure 15: Partitions computed from the pignistic probabilities obtained with ECM (left) and CECM (right).

process the image of a pathological brain in order to detect a tumor. The mere application of the ECM algorithm did not lead to a satisfactory solution, as several parts of the image are associated with a high degree of indetermination. However, introducing a few constraints made it possible to clear up the ambiguity between tumoral and healthy cells and to provide an accurate segmentation of the image.

This research may be extended in several directions. Some authors [18] proposed to add soft constraints rather than crisp ones. A soft constraint may be seen as a relationship between two objects, accompanied with a degree of certainty that this relationship holds. The interest of adding such constraints is twofold. First, one may hope to reduce the negative effect of a small set of constraints on the accuracy of the clustering. Furthermore, the problem of the consistency between the constraints themselves may be tackled to some extent. Besides, the interest of our approach may be illustrated on real-world applications where background knowledge may be provided by experts. In particular, our active-learning scheme could be applied to medical image segmentation. Indeed, a physician may easily label parts of an image as homogeneous regions, or instead require that two regions be classified into different classes.

Acknowledgements

The authors wish to express their thanks to Prof. Catherine Adamsbaum (Hôpital St Vincent de Paul, Paris, France) and Prof. Isabelle Bloch (École Nationale Supérieure des Télécommunications, Paris, France) for providing the brain images.

References

[1] S. Basu, A. Banerjee, and R. Mooney. Semi-supervised clustering by seeding. In *Proceedings of the 19th International Conference on Machine Learning*, pages 19–26, 2002.

[2] S. Basu, A. Banerjee, and R.J. Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the SIAM International Conference on Data Mining*, pages 333–344, 2004.

[3] S. Basu, M. Bilenko, and R.J. Mooney. A probabilist framework for semi-supervised clustering. In *Proceedings of the ACM SIGKDD International Conference on knowledge discovery and data mining*, pages 59–68, 2004.

[4] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, 1981.

[5] I. Bloch. Defining belief functions using mathematical morphology: Application to image fusion under imprecision. *International Journal of Approximate Reasoning*, 48:437–465, 2008.

[6] DA Cohn, Z. Ghahramani, and MI Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[7] R.N. Davé. Clustering relational data containing noise and outliers. *Pattern Recognition Letters*, 12:657–664, 1991.

[8] I. Davidson and S. S. Ravi. Clustering with constraints: Feasibility issues and the k-means algorithm. In *Proceedings of the Fifth SIAM International Conference on Data Mining*, page 138. Society for Industrial Mathematics, 2005.

[9] I. Davidson, K. L. Wagstaff, and S. Basu. Measuring constraints-set utility for partitional clustering algorithms. In *Proceedings of the 10th European Conference on Principle and Practice of Knowledge Discovery in Databases*, pages 115–126, 2006.

[10] T. Denœux and M.-H. Masson. EVCLUS: evidential clustering of proximity data. *IEEE Trans. Systems, Man and Cybernetics: B*, 34:95–109, 2004.

[11] D. Gondek and T. Hofmann. Non-redundant data clustering. *Knowledge and Information Systems*, 12(1):1–24, 2007.

[12] N. Grira, M. Crucianu, and N. Boujemaa. Active semi-supervised fuzzy clustering. *Pattern Recognition*, 41(5):1851–1861, 2008.

[13] D.E. Gustafson and W.C. Kessel. Fuzzy clustering with a fuzzy covariance matrix. In *IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, volume 17, pages 761–765, 1978.

[14] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy cluster analysis: methods for classification, data analysis and image recognition.* John Wiley and Sons, 1999.

[15] D. Klein, S.D. Kamvar, and C.D. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Machine Learning - International Workshop* -, pages 307–314, 2002.

[16] T. Kohonen. *Self-organizing Maps.* Springer, Berlin, 1997.

[17] R. Krishnapuram and J.M. Keller. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(2):98–110, 1993.

[18] M.H.C. Law, A. Topchy, and A.K. Jain. Clustering with soft and group constraints. *Lecture notes in computer science*, 31:662–670, 2004.

[19] Y. Liu, R. Jin, and A.K. Jain. Boostcluster: boosting clustering by pairwise constraints. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 450–459, 2007.

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

[21] M.-H. Masson and T. Denœux. RECM: Relational evidential c-means algorithm. *Pattern Recognition Letters*, 30:1015–1026, 2009.

[22] S. Sen and R.N. Davé. Clustering of relational data containing noise and outliers. In *Fuzzy Systems Proceedings*, volume 2, pages 98–110, 1998.

[23] G. Shafer. *A mathematical theory of evidence.* Princeton university press, Princeton, NJ, 1976.

[24] P. Smets. The transferable belief model for quantified belief representation. In *Hanbook of Defeasible Reasoning and Uncertainty Management Systems*, volume 1, pages 267–301. Kluwer Academic Publishers, 1998.

[25] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

[26] K. Wagstaff. Value, cost, and sharing: Open issues in constrained clustering. *Lecture Notes in Computer Science*, 4747:1, 2007.

[27] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 577–584, 2001.

[28] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, 2003.

[29] R.R. Yager. On the normalization of fuzzy belief structures. *International Journal of Approximate Reasoning*, 14(2-3):127–153, 1996.

[30] Y. Ye and E. Tse. An extension of karmarkar's projective algorithm for convex quadratic programming. *Mathematical Programming*, 44(1):157–179, 1989.

[31] S. Zhong and J. Ghosh. Scalable, balanced model-based clustering. In *Proc. 3rd SIAM Int. Conf. Data Mining*, pages 71–82, 2003.