Evidential clustering with constraints and adaptive metric

Violaine Antoine, Benjamin Quost, Thierry Denœux
Université de Technologie de Compiègne
Heudiasyc, UMR CNRS 6599
BP 20529 60205 Compiègne, France
Email: violaine.antoine@hds.utc.fr

Marie-Hélène Masson
Université de Picardie Jules Verne
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 enables to generate meaningful representations of the data and to improve robustness with respect to outliers. However, in some applications, some kind of background knowledge about the objects or about the clusters is available. To integrate this auxiliary information, constraint-based methods (or semi-supervised) have been proposed. A popular type of constraints specifies whether two objects are in the same (must-link) or in different clusters (cannot-link). 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. We also propose to introduce in the algorithm an adaptive metric so that the constraints may be more easily satisfied. The paper ends with some experimental results.**

Keywords: Constrained clustering; Pairwise constraints; Belief functions theory; Adaptive metric.

I. Introduction

Clustering methods aim at grouping a set of objects into clusters according only to the notion of similarity between their descriptors. However, there exists a wide range of domains where some background knowledge about the problem is available. Adding this extra-information in a clustering algorithm can be useful to guide the method towards a desired solution and to improve the classification accuracy. Wagstaff [15] proposed to introduce in the partition constraints of two types: the first one specifies that two objects have to be in the same cluster (*must-link* constraint) and the second one specifies that two objects must 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 k-means or the fuzzy c-means (FCM), and have become recently a topic of great interest [1], [5], [7], [10], [14], [16]. The FCM algorithm is a method where each object may belong to one or more clusters with different degrees of membership in [0;1]. These degrees of membership are stored into a fuzzy partition matrix $U = (u_{ik})$ and are calculated by minimizing a suitable objective function subject to the constraints $\sum_k u_{ik} = 1 \ \forall i$. Recently, alternative algorithms using the theoretical framework of belief functions have been proposed [6], [11], [12]. 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 $\Omega = \{\omega_1, ..., \omega_c\}$.

One of the algorithms designed to derive a credal partition from data, called ECM, can be considered as a direct extension of FCM. In this paper, we propose to add pairwise constraints in the ECM algorithm, in order to create a new algorithm, called CECM, that will combine the advantages of adding background knowledge and using belief functions. Furthermore, we present a formulation of ECM which adapts the metric using a Mahalanobis distance so that the constraints may be more easily satisfied. The rest of this paper is organized as follows. First, the main fuzzy partitioning algorithms from which ECM is derived are presented in section II. Then the notion of credal partition and the way to derive it from data are described. Section III presents the algorithm CECM. First, we show how to translate in a natural way the available information in terms of constraints on belief masses and how to integrate these constraints in the search of the credal partition. We also describe a version of CECM with an adaptive metric. The last part describes the experimental setting and the results. Finally, some perspectives of our work are presented in a conclusion.

II. Background

A. 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 denotes a matrix composed of the cluster centroids, and $U = (u_{ik})$ defines a fuzzy partition matrix i.e. the degree of membership for each object and each cluster. The FCM algorithm [2] looks for V and U by minimizing the following objective function:

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

In the objective function (1), 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 introducing a Lagrange multiplier with respect to the constraints and setting the partial derivatives of the Lagrangian with respect to the parameters to zero [2]. 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 [4] 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 δ to all objects. The parameter δ controls the amount of data considered as outliers. The membership u_{i*} of an object i to the noise cluster is then:

$$u_{i*} = 1 - \sum_{k=1}^{c} u_{ik} \quad i = 1, n, \tag{2}$$

The objective function to be minimized can be written as:

$$J_{\text{NC}}(U, V) = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{\beta} d_{ij}^2 + \sum_{i=n}^{c} \delta^2 u_{i*}^{\beta}. \tag{3}$$

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

Gustafson and Kessel algorithm 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_l defined as its fuzzy covariance matrix:

$$S_k = \frac{\sum_{i=1}^{n} u_{ik}^{\beta}(\mathbf{x}_i - \mathbf{v}_k)(\mathbf{x}_i - \mathbf{v}_k)^\top}{\sum_{i=1}^{n} u_{ik}^{\beta}} \quad k = 1, c \quad i = 1, n, \tag{4}$$

the distance between an object $\mathbf{x}_i$ and a center $\mathbf{v}_k$ is equal to:

$$d_{ik}^2 = \det(S_k)^{\frac{1}{p}}(\mathbf{x}_i - \mathbf{v}_k)^\top S_k^{-1}(\mathbf{x}_i - \mathbf{v}_k). \tag{5}$$

The adaptation formulas for the memberships and the centers obtained for FCM remain valid as they do not depend on the metric.

B. ECM algorithm [11]

Recently, Masson and Denœux proposed a credibilistic version of Davé's algorithm [11] by replacing the fuzzy partition matrix U with a new partition M called credal partition. The goal is to enhance the concept of hard, fuzzy and possibilistic partition. In this framework, the partial knowledge regarding the class membership of an object is represented by a mass function on the set Ω of possible classes. Thus, beliefs may be given to any subset A of Ω (a 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 it is 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 I. The class of the first object is precisely known whereas the class of the second object is fully unknown. The third object corresponds to a partial knowledge and more particularly to a Bayesian knowledge and the last object, with the totality of the unit mass allocated to the empty set, is considered as an outlier.

Table I
EXAMPLE OF 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 similar to the one of Davé 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 composing A_j:

$$\overline{\mathbf{v}}_j = \frac{1}{|A_j|}\sum_{l=1}^{c} s_{lj}\mathbf{v}_l \quad \text{with} \quad s_{lj} = \begin{cases} 1 & \text{if } \omega_l \in A_j, \\ 0 & \text{else.} \end{cases} \tag{6}$$

The distance d_{ij}^2 between $\mathbf{x}_i$ and the focal set A_j may then be defined by:

$$d_{ij}^2 = ||\mathbf{x}_i - \overline{\mathbf{v}}_j||^2. \tag{7}$$

The ECM algorithm looks for the M and V matrices by minimizing a criteria similar to Davé:

$$J_{\text{ECM}}(M, V) = \sum_{i=1}^{n}\sum_{A_k \neq \emptyset} |A_k|^{\alpha} m_{ik}^{\beta} d_{ik}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta}, \tag{8}$$

subject to:

$$\sum_{k/A_k \subseteq \Omega, A_k \neq \emptyset} m_{ik} + m_{i\emptyset} = 1 \quad \forall i = 1, n, \tag{9}$$

where $m_{i\emptyset}$ denotes the mass of the object $\mathbf{x}_i$ allocated to the empty set. This set is considered as a noise or as an outlier cluster so it is dealt separately. The δ parameter denotes the distance between all the objects and the empty set. An additional weighting coefficient $|A_k|^\alpha$ is introduced to penalize the subsets in Ω with high cardinality, the exponent α allowing to control the degree of penalization. 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.

C. Interpreting a credal partition

As underlined in [11], a credal partition is a rich representation that carries a lot of information about the data. In [11], 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.

III. ECM WITH CONSTRAINTS

A. Class membership of pairs of objects

Let us consider two objects $\mathbf{x}_i$ and $\mathbf{x}_j$ associated with two mass functions m_i and m_j. If we place ourselves in the Cartesian product $\Omega^2 = \Omega\times\Omega$, it is possible to compute a mass function concerning the class membership of both objects. This mass function, denoted $m_{i\times j}$, is the combination of the vacuous extensions of m_i and m_j [13]. It can be written as:

$$m_{i\times j}(A\times B)=m_i(A)\ m_j(B) \quad A,B\subseteq\Omega, A=B\neq\emptyset, \tag{10}$$

$$m_{i\times j}(\emptyset)=m_i(\emptyset)+m_j(\emptyset)-m_i(\emptyset)\ m_j(\emptyset). \tag{11}$$

From $m_{i\times j}$, we can calculate 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 of $\Omega^2 : \theta = \{(\omega_1,\omega_1),(\omega_2,\omega_2),\dots(\omega_c,\omega_c)\}$ and the event "Objects $\mathbf{x}_i$ and $\mathbf{x}_j$ do not belong to the same class" corresponds to its complementary $\overline{\theta}$ in Ω^2. The corresponding plausibilities are the following:

$$pl_{i\times j}(\theta) = \sum_{\{A\times B\subseteq\Omega^2 \ |\ (A\times B)\cap\theta\neq\emptyset\}} m_{i\times j}(A\times B) \tag{12}$$

$$= \sum_{A\cap B\neq\emptyset} m_i(A)\ m_j(B), \tag{13}$$

$$pl_{i\times j}(\overline{\theta}) = 1-m_{i\times j}(\emptyset)-bel_{i\times j}(\theta) \tag{14}$$

$$= 1-m_{i\times j}(\emptyset)-\sum_{k=1}^{c} m_i(\{\omega_k\})\ m_j(\{\omega_k\}). \tag{15}$$

Example 2: Let us consider a new collection of four objects that need to be classified into two classes. A credal partition is given in Table II. Table III gives the masses of the joint membership of object $\mathbf{x}_1$ with the three other objects. The associated plausibilities of θ and $\overline{\theta}$ are given in Table IV.

Table II
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

Table III
MASSES OF JOINT MEMBERSHIP

$F=A\times B$	$m_{1\times2}(F)$	$m_{1\times3}(F)$	$m_{1\times4}(F)$
$\{\omega_1\}\times\{\omega_1\}$	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 IV
PLAUSIBILITIES FOR THE EVENTS θ AND $\overline{\theta}$

F	$pl_{1\times2}(F)$	$pl_{1\times3}(F)$	$pl_{1\times4}(F)$
θ	1	0	1
$\overline{\theta}$	0	1	1

What can be seen thanks to this simple example is that the relevant information in Table IV is contained in the zeros. In fact, nothing can be said about the joint membership of object $\mathbf{x}_1$ and $\mathbf{x}_4$: the plausibilities of θ and $\bar{\theta}$ are equal to 1. On the contrary, the fact that the plausibility of $\bar{\theta}$ is null for $(\mathbf{x}_1,\mathbf{x}_2)$ indicates that these two objects are in the same cluster with certainty. In the same way, the null value of the plausibility of θ for the pair $(\mathbf{x}_1,\mathbf{x}_3)$ indicates with certainty that $\mathbf{x}_1$ and $\mathbf{x}_2$ do not belong to the same class. These remarks are used in the next section to propose a new formulation of ECM integrating pairwise constraints (must-link or cannot-link).

B. Objective function of CECM

Let us consider now 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 expressed as 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 "compatible" with the similarities computed from the data and the constraints. A natural requirement is that $pl_{i\times j}(\theta)$ must be as low as possible if $(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}$. In the same way, $(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}$ 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:

$$J_{\text{CECM}}(M,V) = J_{\text{ECM}}(M,V) + \gamma \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}} pl_{i\times j}(\overline{\theta}) + \eta \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}} pl_{i\times j}(\theta), \quad (16)$$

such that the constraints (9) are respected. The coefficients γ and η control the tradeoff between the objective function of ECM and the constraints.

C. Adaptive metric

In the ECM algorithm, the distance d_{ik}^2 between the object $\mathbf{x}_i$ and the centroid $\overline{\mathbf{v}}_k$ is an 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 an Euclidean model. To modify the previous algorithm, we follow an approach inspired from Gustafson and Kessel algorithm [8] and well described in [9]. Let S_l denote a $(p \times p)$ matrix associated to cluster ω_l $(l = 1, c)$ inducing a norm $||\mathbf{x}||^2_{S_l} = \mathbf{x}^\top S_l \mathbf{x}$. Using the same approach that we used for the centroids, we define the matrix $\overline{S}_j$ associated to a non singleton A_j by averaging the matrices associated with the classes composing A_j:

$$\overline{S}_j = \frac{1}{|A_j|}\sum_{l=1}^{c} s_{lj} S_l \quad \forall A_j \subseteq \Omega, A_j \neq \emptyset. \quad (17)$$

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

$$d_{ij}^2 = ||\mathbf{x}_i - \overline{\mathbf{v}}_j||^2_{\overline{S}_j} = (\mathbf{x}_i - \overline{\mathbf{v}}_j)^\top \overline{S}_j (\mathbf{x}_i - \overline{\mathbf{v}}_j). \quad (18)$$

D. Optimization

We propose an alternate optimization scheme in order to fix the partition matrix M, the centroids matrix V, and the metrics S_l.

1) Optimization with respect to the masses: With regard to the belief masses, in the general case, the problem is complex and the derivation from the optimality conditions of a direct update equation of the m_{ij} like in ECM is no longer possible. However, if we fix $\beta = 2$ and using relation (14) then the objective function (16) becomes:

$$J_{\text{CECM}}(M,V,S_1,\cdots,S_c) = \sum_{i=1}^{n}\sum_{A_k\neq\emptyset} |A_k|^\alpha m_{ik}^2 d_{ik}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^2 - \gamma \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}} m_{i\times j}(\emptyset) - \gamma \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}} bel_{i\times j}(\theta) + \eta \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}} pl_{i\times j}(\theta) + \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}} \gamma. \quad (19)$$

Note that the last term of equation (19), 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}. As we have linear constraints, a classical optimization method [17] can be used and the convergence is insured in a reasonable time, whatever the number of constraints considered. Note that, due to the quadratic optimization problem, our approach remains limited to a few hundred of samples and a small number of classes ($c \leq 5$).

2) Optimization with respect to the cluster centers: We 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:

$$\frac{\partial J_{\text{CECM}}}{\partial \mathbf{v}_l} = \sum_{i=1}^{n}\sum_{A_j\neq\emptyset} |A_j|^\alpha m_{ij}^2 \frac{\partial d_{ij}^2}{\partial \mathbf{v}_l} \quad l = 1, c. \quad (20)$$

$$\frac{\partial d_{ij}^2}{\partial \mathbf{v}_l} = 2(s_{lj})\overline{S}_j(\mathbf{x}_i - \bar{\mathbf{v}}_j)(-\frac{1}{|A_j|}) \quad l = 1, c. \quad (21)$$

From (20) and (21) we thus have:

$$\frac{\partial J_{\text{CECM}}}{\partial \mathbf{v}_l} = -2\sum_{i=1}^{n}\sum_{A_j\neq\emptyset} |A_j|^{\alpha-1} m_{ij}^2 s_{lj} \overline{S}_j (\mathbf{x}_i - \frac{1}{|A_j|}\sum_k s_{kj}\mathbf{v}_k). \quad (22)$$

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

$$\sum_i \sum_{A_j\ni\omega_l} |A_j|^{\alpha-1} m_{ij}^2 \overline{S}_j \mathbf{x}_i = \sum_k \sum_i \sum_{A_j\supseteq\{\omega_k,\omega_l\}} |A_j|^{\alpha-2} m_{ij}^2 \overline{S}_j \mathbf{v}_k \quad l = 1, c. \quad (23)$$

Let $\mathbf{F}^{(l,i)}$ and $\mathbf{G}^{(l,k)}$ denote two $(p \times p)$ matrices:

$$\mathbf{F}^{(l,i)} = \sum_{A_j\ni\omega_l} |A_j|^{\alpha-1} m_{ij}^2 \overline{S}_j \quad l = 1, c \quad i = 1, n, \quad (24)$$

$$\mathbf{G}^{(l,k)} = \sum_i \sum_{A_j\supseteq\{\omega_k,\omega_l\}} |A_j|^{\alpha-2} m_{ij}^2 \overline{S}_j \quad k, l = 1, c. \quad (25)$$

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:

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}^{(1,1)} & \mathbf{F}^{(1,2)} & \cdots & \mathbf{F}^{(1,n)} \\ \mathbf{F}^{(2,1)} & \mathbf{F}^{(2,2)} & \cdots & \mathbf{F}^{(2,n)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{F}^{(c,1)} & \mathbf{F}^{(c,2)} & \cdots & \mathbf{F}^{(c,n)} \end{pmatrix} \quad (26)$$

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}^{(1,1)} & \mathbf{G}^{(1,2)} & \cdots & \mathbf{G}^{(1,c)} \\ \mathbf{G}^{(2,1)} & \mathbf{G}^{(2,2)} & \cdots & \mathbf{G}^{(2,c)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}^{(c,1)} & \mathbf{G}^{(c,2)} & \cdots & \mathbf{G}^{(c,c)} \end{pmatrix} \quad (27)$$

Let us stack all objects $\mathbf{x}_i$ 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:

$$\mathbf{G}V = \mathbf{F}\mathbf{X}. \tag{28}$$

3) Optimization with respect to the metrics S_l: We now consider that M and V are fixed. We follow the same line of reasoning than Gustafson and Kessel. In order to avoid a minimization of J_{CECM} by matrices S_l with zeros entries, we impose a constant volume of the clusters using the constraints $\det(S_l) = 1 \ \forall l = 1, c$. 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:

$$\mathcal{L} = J_{\text{CECM}}(M, V, S_1, \cdots, S_c) - \sum_{k=1}^{c} \lambda_k \left(\det(S_k) - 1\right) \tag{29}$$

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

$$\frac{\partial \mathcal{L}}{\partial S_l} = \sum_i \sum_{A_j \neq \emptyset} m_{ij}^2 |A_j|^{\alpha-1} s_{lj} (\mathbf{x}_i - \overline{\mathbf{v}}_j)(\mathbf{x}_i - \overline{\mathbf{v}}_j)^\top - \lambda_l \det(S_l) S_l^{-1}. \tag{30}$$

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

$$\Sigma_l = \sum_i \sum_{A_j \ni \omega_l} m_{ij}^2 |A_j|^{\alpha-1} (\mathbf{x}_i - \overline{\mathbf{v}}_j)(\mathbf{x}_i - \overline{\mathbf{v}}_j)^\top \quad l = 1, c. \tag{31}$$

Note that Σ_l can be considered as the analog in the evidential framework of the fuzzy covariance matrix. From (30), we have:

$$\Sigma_l = \lambda_l S_l^{-1} \quad l = 1, c, \tag{32}$$

$$\Rightarrow \quad \Sigma_l S_l = \lambda_l I \quad l = 1, c, \tag{33}$$

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

$$\det(\Sigma_l S_l) = \det(\Sigma_l)\det(S_l) = \det(\Sigma_l) = \lambda_l^p \quad l = 1, c. \tag{34}$$

It follows that

$$\lambda_l = \det(\Sigma_l)^{\frac{1}{p}} \quad l = 1, c. \tag{35}$$

If we replace λ_l by its expression and using (32), we finally obtain:

$$S_l = \det(\Sigma_l)^{\frac{1}{p}} \Sigma_l^{-1} \quad l = 1, c. \tag{36}$$

IV. Experimental results

A. Toy data set

In order to illustrate the interest of introducing constraints, we created a synthetic dataset. It consists of two classes of patterns in a two-dimensional space. In each class, 100 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, and with a common covariance. The standard version of ECM algorithm was run using an Euclidean distance, with $c = 2$, $\rho^2 = 100$ and $\alpha = 1$. The credal partition obtained shows a diagonal boundary between the two classes. 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 1 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. The add of a few number of constraints randomly chosen enables 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 2.

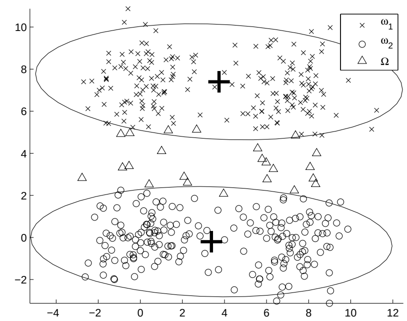

Figure 1. Hard credal partition obtained with ECM (no constraint) using an adaptive metric.

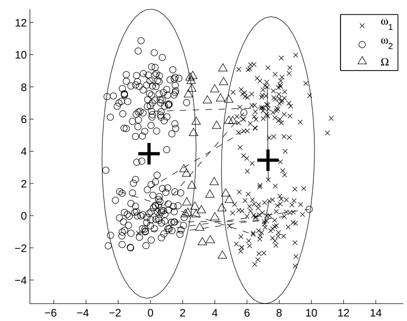

Figure 2. CECM with a Mahalanobis distance and 10 constraints; plain lines represent *must-link* constraints, and dashed lines *cannot-link* constraints.

B. Medical image segmentation

The interest of CECM will now be illustrated using an example in medical imaging taken from [3]. An image of a pathological brain was acquired using magnetic resonance imaging. It is represented in Figure 3. In this image, according to the gray levels of the pixels, three main areas may be distinguished: the brightest area corresponds to a tumor, 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. Starting from the gray levels

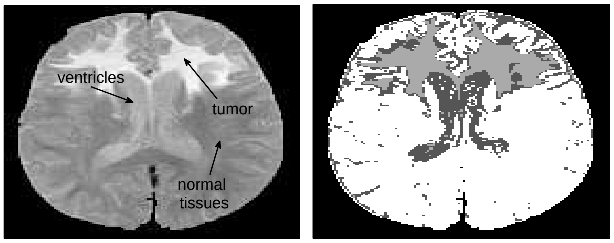

Figure 3. (Left): original image; (right): hard credal partition obtained from ECM with an Euclidean metric (white: ω_1, light gray: ω_2, dark gray: Ω).

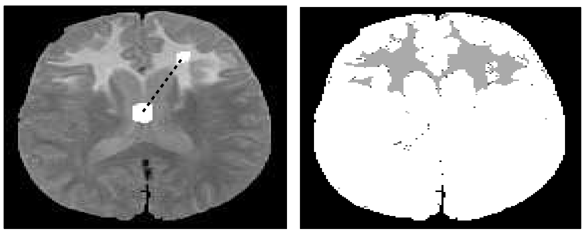

Figure 4. (Left): must-link (white areas) and cannot-link constraints (dashed line); (Right): hard credal partition obtained from CECM with an adaptive metric (white: ω_1, light gray: ω_2, dark gray: Ω).

of the pixels (rescaled between 0 and 1), ECM, with $c = 2$, $\alpha = 2$, and $\rho^2 = 10$, finds a hard credal partition represented in Figure 3. 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 4. 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 4. It may be seen that the constraints made possible to raise the indetermination concerning the pixels allocated to Ω and thus to properly isolate the tumor. As a matter of comparison, the partitions computed from the pignistic probabilities obtained by ECM and CECM are given in Figure 5.

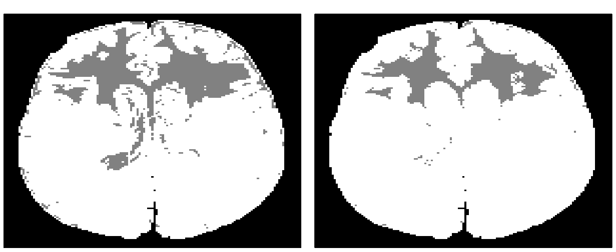

Figure 5. Partitions computed from the pignistic probabilities; (Left): Results of ECM; (Right) results of CECM.

V. Conclusion

In this paper, we have presented a new clustering method called CECM based on the belief functions theory. It is an extension of the evidential clustering algorithm ECM. The contribution of the paper is twofold. First, we have proposed to add pairwise constraints. Second, we have introduced an adaptive metric in the algorithm. This distance, more general than the Euclidean distance, treats non spherical classes and adjusts to the add of constraints. Experiments have shown that these two extensions make it possible to guide the algorithm towards desired solutions.

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, 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.
[2] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, 1981.
[3] I. Bloch. Defining belief functions using mathematical morphology : Application to image fusion under imprecision. *International Journal on Approximate Reasoning*, 48(2):437–465, 2008.
[4] R.N. Davé. Clustering relational data containing noise and outliers. *Pattern Recognition Letters*, 12:657–664, 1991.
[5] 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.
[6] T. Denœux and M.-H. Masson. EVCLUS: evidential clustering of proximity data. *IEEE Trans. Systems, Man and Cybernetics: B*, 34:95–109, 2004.
[7] N. Grira, M. Crucianu, and N. Boujemaa. Active semi-supervised fuzzy clustering. *Pattern Recognition*, 41(5):1851–1861, 2008.
[8] 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.
[9] 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.
[10] 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.
[11] M.-H. Masson and T. Denœux. ECM: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41:1384–1397, 2008.
[12] M.-H. Masson and T. Denœux. RECM: Relational evidential c-means algorithm. *Pattern Recognition Letters*, 30:1015–1026, 2009.
[13] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.
[14] K. Wagstaff. Value, cost, and sharing: Open issues in constrained clustering. *Lecture Notes in Computer Science*, 4747:1, 2007.
[15] 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.
[16] E.P. Xing, A.Y. Ng, Jordan M.I., and S.J. 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.
[17] Y. Ye and E. Tse. An extension of karmarkar's projective algorithm for convex quadratic programming. *Mathematical Programming*, 44(1):157–179, 1989.