CECM: Adding Pairwise Constraints To Evidential Clustering

Violaine Antoine, Benjamin Quost, Marie-Hélène Masson, and Thierry Denœux

***Abstract*— Fuzzy or hard partitioning methods aim at grouping objects according to their similarity. Recently, a new concept of partition based on belief function theory, called credal partition, has been proposed and has been shown to generate meaningful description of the data. Hard, fuzzy or credal partitions are generally obtained using unsupervised learning methods, using only the numeric description between two objects to compute their similarity. 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 (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). We propose here a new algorithm, called CECM, which computes a credal partition using a constrained clustering method. We show how to translate the available information into constraints, and how to integrate them in the search of the credal partition. The paper ends with some experimental results. Results of CECM are compared to other constrained clustering algorithms. Then an application in image segmentation is described.**

I. Introduction

Clustering is a classical data analysis method that aims at grouping a set of objects into clusters. Classically, clustering proceeds from unsupervised learning: indeed, the clusters are based on the similarity between the descriptors of the objects only. However, there are some situations in which some background knowledge about the problem is available. This extra-information may be used to guide the clustering algorithm towards a desired solution, and thus 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 [1], the *model* level with the requirement of balanced clusters [2] or the specification of non desired solutions [3], or at the *instance* level. Wagstaff [4] proposed to introduce two types of instance-level constraints. A *must-link* constraint specifies that two objects have to be in the same cluster; a *cannot-link* constraint, that they should not be put in the same cluster. 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 [5], [6], [7], [1], [8]. They have been incorporated in many different ways, generally by including a penalty term in the objective function [9], [10] or by altering the distances between objects with respect to the constraints [11], [5].

Violaine Antoine, Benjamin Quost, and Thierry Denœux are members of Heudiasyc Laboratory, University of Technology of Compiègne, Centre de Recherche de Royallieu, BP20529, 60205 Compiègne, France (email: violaine.antoine@hds.utc.fr).

Marie-Hélène Masson is assistant professor at University of Picardie Jules Verne and a member of Heudiasyc Laboratory, BP20529, 60205 Compiègne, France (email: mmasson@hds.utc.fr).

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 subject to the constraints $\sum_k u_{ik} = 1 \; \forall i$. Each number $u_{ik} \in [0, 1]$ is interpreted as the degree of membership of object i to cluster k. The FCM algorithm is known to produce sometimes counterintuitive results, and to have poor robustness against noise and outliers. Therefore, possibilistic methods [12], [13], and more recently algorithms using the theoretical framework of belief functions [14], [15], [16], have been proposed. These latter 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. A credal partition consists in allocating, for each object, a mass of belief 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. The Evidential C-Means (ECM) algorithm [15], that derives a credal partition from data, can be considered as a direct extension of FCM.

In this paper, we propose to introduce pairwise constraints in ECM. The resulting algorithm, called CECM, thus combines the advantages of adding background knowledge and using belief functions. Besides, we present a formulation of ECM that adapts the metric using a Mahalanobis distance, so that the constraints may be more easily satisfied. The remaining of this paper is organized as follows. In Section II, the main fuzzy partitioning algorithms from which ECM is derived are presented. Then, a brief overview of the theory of belief functions is provided, and particularly the notion of credal partition. Section III introduces the CECM algorithm. 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 IV, we also describe a version of CECM allowing to automatically modify the metric according to the constraints. Section V describes some experiments. Several results are presented. We first compare the performances of CECM with those of other constrained clustering algorithms. We also demonstrate the usefulness of the method with an an application in image segmentation. Finally, section VI concludes this paper.

978-1-4244-8126-2/10/$26.00

II. Background

A. Fuzzy C-Means And Variants

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set 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 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 [17] computes V and U so as to minimize the following objective function:

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

In the objective function (1), d_{ik} represents the Euclidean distance between the object $\mathbf{x}_i$ and the centroid $\mathbf{v}_k$. Parameter $\beta > 1$ 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 of the parameters are obtained by computing the Lagrangian formulation of the optimization problem and writing its Karush-Kuhn-Tucker (KKT) optimality conditions [17]. We obtain:

$$\mathbf{v}_k = \frac{\sum_{i=1}^{n} u_{ik}^{\beta} \mathbf{x}_i}{\sum_{i=1}^{n} u_{ik}^{\beta}} \quad k = 1, c, \tag{2}$$

$$u_{ij} = \frac{d_{ij}^{-2/(\beta-1)}}{\sum_{k=1}^{c} d_{ik}^{-2/(\beta-1)}} \quad i = 1, n \quad j = 1, c. \tag{3}$$

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 [18] proposed a variant of FCM called the "noise-clustering" algorithm (NC). It consists in adding to the c initial clusters a "noise" cluster. A parameter ρ defines the distance of this cluster to the others, and thus controls the amount of data considered as outliers. The membership u_{i*} of an object i to the noise cluster is given by:

$$u_{i*} = 1 - \sum_{k=1}^{c} u_{ik} \quad i = 1, n, \tag{4}$$

The objective function to be minimized thus becomes:

$$J_{\text{NC}}(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{\beta} d_{ij}^2 + \sum_{i=n}^{c} \rho^2 u_{i*}^{\beta}. \tag{5}$$

As in FCM, writing the KKT conditions of the Lagrangian leads to direct adaptation formulas for the memberships and the cluster centers.

The Gustafson and Kessel algorithm [19] extends FCM by using an adaptive distance. Thus, clusters of different geometrical shapes may be detected. Each cluster has its own norm-inducing matrix S_k defined as its fuzzy covariance matrix:

$$S_k = \frac{\sum_{i=1}^{n} u_{ik}^{\beta} (\mathbf{x}_i - \mathbf{v}_k)(\mathbf{x}_i - \mathbf{v}_k)^t}{\sum_{i=1}^{n} u_{ik}^{\beta}} \quad k = 1, c \quad i = 1, n. \tag{6}$$

The distance between an object $\mathbf{x}_i$ and a center $\mathbf{v}_k$ is then:

$$d_{ik}^2 = \det(S_k)^{\frac{1}{p}} (\mathbf{x}_i - \mathbf{v}_k)^t S_k^{-1} (\mathbf{x}_i - \mathbf{v}_k). \tag{7}$$

Equation (6) 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.

B. Belief Functions

The Dempster-Shafer theory of evidence [20], [21] (or belief function theory) is a theoretical framework for representing partial and unreliable information. In this section, only the main concepts are recalled.

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

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{8}$$

Any subset $A \subseteq \Omega$ such that $m(A) > 0$ is called a focal set of m. The quantity $m(A)$ 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. A bba m expresses total ignorance if $m(\Omega) = 1$, and full certainty whenever $m(A) = 1$ for some $A \subseteq \Omega$ (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. Otherwise, $m(\emptyset)$ may be interpreted as the belief that the actual value of ω does not belong to Ω [22].

Given a bba m, the plausibility function $pl : 2^{\Omega} \rightarrow [0, 1]$ and the belief function $bel : 2^{\Omega} \rightarrow [0, 1]$ are defined by:

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \quad \forall A \subseteq \Omega, \tag{9}$$

$$bel(A) = \sum_{B \subseteq A, B \neq \emptyset} m(B) \quad \forall A \subseteq \Omega. \tag{10}$$

Functions bel and pl are linked by the following relation:

$$pl(A) = 1 - m(\emptyset) - bel(\overline{A}), \tag{11}$$

where $\overline{A}$ denotes the complement of A. The quantity $bel(A)$ is interpreted as a degree of justified support given to A by the available evidence, and $pl(A)$ measures to what extent one fails to believe in hypotheses incompatible with A.

In order to make a decision regarding the value of ω, a bba m may be transformed into a pignistic probability distribution [21] $BetP$. For a normal bba, we have:

$$BetP(\omega) = \sum_{\omega \in A} \frac{m(A)}{|A|} \quad \forall \omega \in \Omega, \tag{12}$$

where $|A|$ denotes the cardinality of $A \subseteq \Omega$. If $m(\emptyset) \neq 0$, a normalization step must precede 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 Ω [23].

C. ECM Algorithm

Masson and Denœux proposed a credibilistic version of Davé's algorithm [15], where the fuzzy partition matrix U is replaced with a more general kind of partition M, called a credal partition. Partial knowledge regarding the class membership of an object i is represented by a bba m_i on the set Ω of possible classes. Thus, any subset A of Ω may receive support. 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 I. 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, what is represented by allocating the whole unit mass to the empty set.

TABLE I
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, which corresponds to a situation of complete knowledge, then M defines a conventional, crisp partition of the set of objects. If all the m_i are Bayesian, M specifies a fuzzy partition. With focal elements being singletons of Ω or the empty set, a partition with a noise cluster as in the NC algorithm is recovered.

The ECM algorithm derives 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 computing, 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. 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:

$$\overline{\mathbf{v}}_j = \frac{1}{|A_j|} \sum_{l=1}^{c} s_{lj} \mathbf{v}_l, \tag{13}$$

with

$$s_{lj} = \begin{cases} 1 & \text{if } \omega_l \in A_j, \\ 0 & \text{else.} \end{cases} \tag{14}$$

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

$$d_{ij} = ||\mathbf{x}_i - \overline{\mathbf{v}}_j||. \tag{15}$$

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

$$J_{\text{ECM}}(M,V) = \sum_{i=1}^{n} \sum_{A_k \neq \emptyset} |A_k|^{\alpha} m_{ik}^{\beta} d_{ik}^2 + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta}, \tag{16}$$

subject to:

$$\sum_{k/A_k \subseteq \Omega, A_k \neq \emptyset} m_{ik} + m_{i\emptyset} = 1 \quad \forall i = 1,n, \tag{17}$$

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 iterative optimization with the alternate update of the masses and the centroids. The KKT conditions gives the following adaptation rule for the masses: for $i = 1, \ldots, n$ and $A_j \neq \emptyset$,

$$m_{ij} = \frac{|A_j|^{-\alpha/(\beta-1)} d_{ij}^{-2/(\beta-1)}}{\sum_{A_k \neq \emptyset} |A_k|^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)} + \rho^{-2/(\beta-1)}} \tag{18}$$

and

$$m_{i\emptyset} = 1 - \sum_{A_j \neq \emptyset} m_{ij} \quad \forall i = 1,n. \tag{19}$$

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 for $l = 1, \ldots, c$ and $q = 1, \ldots, p$ by:

$$\mathbf{B}_{lq} = \sum_{i=1}^{n} x_{iq} \sum_{A_j \ni \omega_l} |A_j|^{\alpha-1} m_{ij}^{\beta} \tag{20}$$

and $\mathbf{H}$ a matrix of size $(c \times c)$ given $(k, l = 1, \ldots, c)$ by:

$$\mathbf{H}_{lk} = \sum_{i} \sum_{A_j \supseteq \{\omega_k, \omega_l\}} |A_j|^{\alpha-2} m_{ij}^{\beta}. \tag{21}$$

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

$$\mathbf{H}V = \mathbf{B}, \tag{22}$$

which can be solved using a standard linear system solver. Details on the calculation of Equations (18) to (22) are provided in [15]. Note that, in practice, the resolution of (22) 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 alternates the optimization of M and V until convergence.

D. Interpreting A Credal Partition

As underlined in [15], a credal partition is a rich representation that carries a lot of information about the data. In [15], 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.

Example 2: Let us consider the credal partition presented in Table I. The corresponding pignistic probabilities (using Yager's normalization) are given in Table II.

TABLE II
PIGNISTIC PROBABILITIES FOR THE CREDAL PARTITION OF TABLE I

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

III. ECM WITH CONSTRAINTS

A. 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 [21]. As shown in [14], we have, for $A, B \subseteq \Omega$, $A \neq \emptyset$, $B \neq \emptyset$:

$$m_{i\times j}(A \times B) = m_i(A)\ m_j(B), \tag{23}$$

$$m_{i\times j}(\emptyset) = m_i(\emptyset) + m_j(\emptyset) - m_i(\emptyset)\ m_j(\emptyset). \tag{24}$$

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:

$$pl_{i\times j}(\theta) = \sum_{A\cap B\neq\emptyset} m_i(A)\ m_j(B), \tag{25}$$

$$pl_{i\times j}(\overline{\theta}) = 1 - m_{i\times j}(\emptyset) - \sum_{k=1}^{c} m_i(\{\omega_k\})m_j(\{\omega_k\}). \tag{26}$$

Example 3: Let us consider a new collection of four objects to be classified into two classes. A credal partition expressing certain knowledge about the membership of the objects is given in Table III. Table IV gives the mass functions of the joint membership of $\mathbf{x}_1$ with the three other objects. The associated plausibilities $pl(\theta)$ and $pl(\overline{\theta})$ are given in Table V.

TABLE III
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 IV
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 V
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

This simple example shows how the joint membership of two objects may be represented using the plausibilities $pl(\theta)$ and $pl(\overline{\theta})$. In simple terms, the relevant information in Table V 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 $pl_{1\times 2}^{\Omega\times\Omega}(\overline{\theta}) = 0$ indicates that ($\mathbf{x}_1$ and $\mathbf{x}_2$) are certainly in the same cluster. Equivalently, the null value of the plausibility $pl_{1\times 3}^{\Omega\times\Omega}(\theta)$ expresses 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.

B. 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:

$$J_{\text{CECM}}(M,V) = J_{\text{ECM}}(M,V) + \frac{\gamma}{|\mathcal{M}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}} pl_{i\times j}(\overline{\theta}) + \frac{\eta}{|\mathcal{C}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}} pl_{i\times j}(\theta), \quad (27)$$

such that the constraints (17) 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. Note that we choose to express the penalization using plausibilities rather than beliefs since this quantities depends only on the mass given to singletons.

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, note that the two penalty terms added to (27) do not depend on the cluster centroids. The same update scheme for the centroids (Equations (20) to (22)) can thus be used in CECM. Suppose furthermore that we fix $\beta = 2$; using Equation (26), we get:

$$J_{\text{CECM}}(M,V) = \sum_{i=1}^{n}\sum_{A_k\neq\emptyset}|A_k|^{\alpha}m_{ik}^2 d_{ik}^2 + \sum_{i=1}^{n}\rho^2 m_{i\emptyset}^2 - \frac{\gamma}{|\mathcal{M}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}} m_{i\times j}(\emptyset) - \frac{\gamma}{|\mathcal{M}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}} bel_{i\times j}(\theta) + \frac{\eta}{|\mathcal{C}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}} pl_{i\times j}(\theta) + \gamma. \quad (28)$$

Note that the last term of Equation (28), 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 [24] can be used and the convergence is insured in a reasonable time. Note that the complexity of our approach is linear with the respect to the number of samples and is exponential with the respect to the number of classes, so the algorithm remains limited to a few hundreds of samples and a small number of classes ($c \leq 5$). 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
Initialization of V
repeat
1) Calculate the new masses by solving the quadratic programming problem defined by (28) subject to (17).
2) Calculate the new centroids by solving the linear system defined by Equations (20) to (22).
until No significant change in V between two successive iterations

IV. CECM With An Adaptive Metric

A. Model

The use of a Mahalanobis distance, instead of an Euclidean distance like in the ECM algorithm, 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 spherical model. To modify the previous algorithm, we follow an approach inspired from Gustafson and Kessel [19], [25]. 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 as for the centroids, we compute the matrix $\overline{S}_j$ associated with a non singleton A_j by averaging the matrices associated with the classes $\omega_k \in A_j$:

$$\overline{S}_j = \frac{1}{|A_j|}\sum_{l=1}^{c} s_{lj} S_l, \quad \forall A_j \subseteq \Omega, A_j \neq \emptyset. \quad (29)$$

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

$$d_{ij}^2 = ||\mathbf{x}_i - \overline{\mathbf{v}}_j||_{\overline{S}_j}^2 = (\mathbf{x}_i - \overline{\mathbf{v}}_j)^t \overline{S}_j (\mathbf{x}_i - \overline{\mathbf{v}}_j). \quad (30)$$

The new criterion to be minimized thus becomes:

$$J_{\text{CECM}}(M,V,S_1,\cdots,S_c) = \sum_{i=1}^{n}\sum_{A_k\neq\emptyset}|A_k|^{\alpha}m_{ik}^2||\mathbf{x}_i - \overline{\mathbf{v}}_j||_{\overline{S}_j}^2 + \sum_{i=1}^{n}\rho^2 m_{i\emptyset}^2 + \frac{\gamma}{|\mathcal{M}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{M}} pl_{i\times j}(\overline{\theta}) + \frac{\eta}{|\mathcal{C}|}\sum_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{C}} pl_{i\times j}(\theta). \quad (31)$$

Note that the minimization of (31) 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. Unlike in [19], the determination of the centers takes here the metric explicitly into account, as shown below.

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. Computing partial derivatives of J_{CECM} with respect to the centers $\mathbf{v}_k$ gives c update equations for the centers $\mathbf{v}_k$:

$$\sum_i \sum_{A_j\ni\omega_l} |A_j|^{\alpha-1} m_{ij}^2 \overline{S}_j \mathbf{x}_i = \sum_k \sum_i \sum_{A_j\supseteq\{\omega_k,\omega_l\}} |A_j|^{\alpha-2} m_{ij}^2 \overline{S}_j \mathbf{v}_k \quad l = 1, c. \quad (32)$$

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

$$\mathbf{F}^{(l,i)} = \sum_{A_j \ni \omega_l} |A_j|^{\alpha-1} m_{ij}^2 \overline{S}_j \quad l = 1, c \quad i = 1, n, \quad (33)$$

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

$$\mathbf{G}^{(l,k)} = \sum_{i} \sum_{A_j \supseteq \{\omega_k, \omega_l\}} |A_j|^{\alpha-2} m_{ij}^2 \overline{S}_j \quad k, l = 1, c. \quad (34)$$

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:

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}^{(1,1)} & \mathbf{F}^{(1,2)} & \dots & \mathbf{F}^{(1,n)} \\ \mathbf{F}^{(2,1)} & \mathbf{F}^{(2,2)} & \dots & \mathbf{F}^{(2,n)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{F}^{(c,1)} & \mathbf{F}^{(c,2)} & \dots & \mathbf{F}^{(c,n)} \end{pmatrix} \quad (35)$$

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}^{(1,1)} & \mathbf{G}^{(1,2)} & \dots & \mathbf{G}^{(1,c)} \\ \mathbf{G}^{(2,1)} & \mathbf{G}^{(2,2)} & \dots & \mathbf{G}^{(2,c)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}^{(c,1)} & \mathbf{G}^{(c,2)} & \dots & \mathbf{G}^{(c,c)} \end{pmatrix} \quad (36)$$

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:

$$\mathbf{G}V = \mathbf{F}\mathbf{X}. \quad (37)$$

2) Optimization With Respect To The Matrices S_l: We now consider that M and V are fixed. To determine the matrices S_l, we follow the same line of reasoning as Gustafson and Kessel. Imposing that the clusters have a constant volume, using the constraints $\det(S_l) = 1$ for all $l = 1, c$, is necessary to avoid the degenerate solution consisting of matrices S_l with zero entries. 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:

$$\mathcal{L}(S_1, \cdots, S_c, \lambda_1, \cdots, \lambda_c) = J_{\text{CECM}}(M, V) - \sum_{k=1}^{c} \lambda_k \left(\det(S_k) - 1\right) \quad (38)$$

Applying the KKT conditions to this Lagrangian leads to the following update equations for the covariance matrices S_l:

$$S_l = \det(\Sigma_l)^{\frac{1}{p}} \Sigma_l^{-1} \quad l = 1, c, \quad (39)$$

with Σ_l being defined, for $l = 1, \dots, c$, by:

$$\Sigma_l = \sum_{i} \sum_{A_j \ni \omega_l} m_{ij}^2 |A_j|^{\alpha-1} (\mathbf{x}_i - \overline{\mathbf{v}}_j)(\mathbf{x}_i - \overline{\mathbf{v}}_j)^t. \quad (40)$$

Note that Σ_l can be considered as the analog in the evidential framework of the fuzzy covariance matrix. Furthermore, Σ_l is invertible since each $(\mathbf{x}_i - \overline{\mathbf{v}}_j)(\mathbf{x}_i - \overline{\mathbf{v}}_j)^t$ is symmetric, positive and semi-definite; hence, 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 $l = 1, c$
Initialization of V
repeat
1) Calculate the new masses by solving the quadratic programming problem defined by (31) subject to (17).
2) Calculate the new centroids by solving the linear system defined by Equation (32).
3) Calculate the new matrices S_l, $l = 1, c$ using (40) and (39).
until No significant change in V between two successive iterations

V. Experimental Results

A. Methodology

We show here how adding pairwise constraints may improve the classification accuracy. We use two real classical data sets for which a reference partition is known. Various measures may be used to measure the degree to which the predicted class labels match the actual ones. Since the label of each cluster is arbitrary and does not reflect any ground truth (unlike in supervised classification), a practical way is to check whether pairs of points are assigned to the same cluster or not in the predicted and actual partitions. Among the criteria based on this principle, the F-measure is one of the most popular. It is defined as the harmonic mean of two quantities traditionally used in information retrieval, the precision and recall. More precisely the pairwise F-measure is defined as:

$$\text{F-measure} = \frac{2 \times \text{Prec} \times \text{Recall}}{\text{Prec+Recall}}, \quad (41)$$

with

$$\text{Prec} = \frac{\text{\# pairs correctly predicted in the same cluster}}{\text{\# total pairs predicted in the same cluster}}, \quad (42)$$

and

$$\text{Recall} = \frac{\text{\# pairs correctly predicted in the same cluster}}{\text{\# total pairs actually in the same cluster}}. \quad (43)$$

The performances of CECM are compared to those of unsupervised clustering algorithms: FCM [17] and K-Means. We also provide comparison with other classical algorithms incorporating pairwise constraints:

- COP-KMeans [4], which introduces constraints in the K-Means algorithm and ensures that they are satisfied at each iteration.

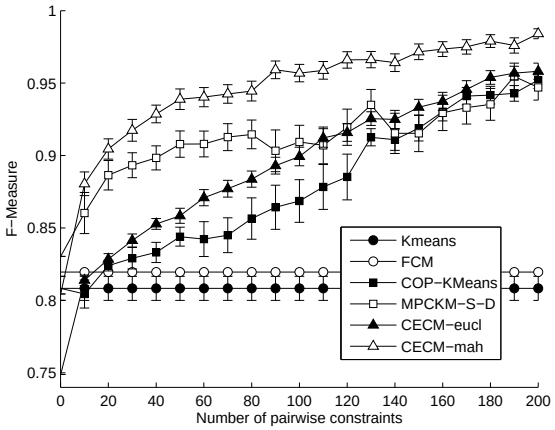

Fig. 1. Results obtained by different clustering algorithms, Iris.

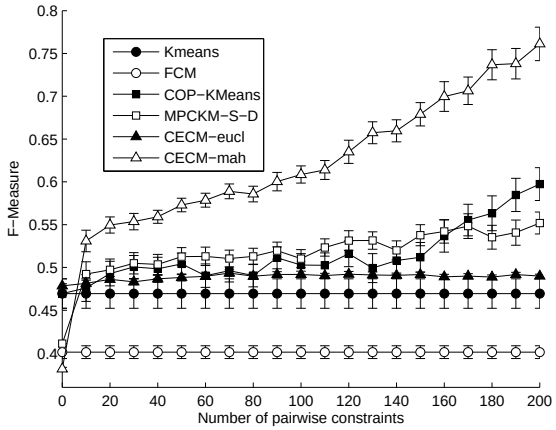

Fig. 2. Results obtained by different clustering algorithms, Letters-IJL.

- MPCKM-S-D clustering [26], which utilizes constraints for seeding the initial clusters and then apply a metric learning scheme by using a single (S) and diagonal (D) parametrized matrix for all the clusters [1].

B. Results On Two Real Data Sets

Experiments were conducted on two data set from the UCI repository: *Iris* and *Letters*. The *Iris* data set is composed of three classes of patterns in a four-dimensional space of 50 data each. For the *Letters* data set, we selected three classes corresponding to the letters $\{I, J, L\}$. Then, following [26], we randomly selected 10% of the data in each class, so that we obtained 227 data in a 16-dimensional space. The three classes are roughly balanced.

We used two clustering schemes for CECM: CECM-eucl, which exploits the Euclidean distance, and the CECM-mah, which introduces the Mahalanobis distance. We used parameter values of $\gamma = \eta = 1$ and $\rho^2 = 1000$ for both schemes and both data sets.

Figures 1 and 2 show the average F-Measure plotted against a varying number of constraints. Averages were computed on 100 experiments using randomly selected constraints. Must-link and Cannot-link constraints are set using the true known classes. Note that noisy or incoherent constraints, especially if they are prevalent in the set of constraints, are likely to deteriorate the solution provided by CECM. The coherence of the constraints should thus be checked before running CECM (see e.g. [27]).

It may be seen that constraint-based approaches outperform unsupervised clustering. Algorithms using a metric learning (MPCKM-S-D and CECM-mah) give the best results. This was predictable since the data sets processed have non-spherical classes. CECM-mah produces better clustering than MPCKM-S-D. Likely, the reason is that the former method uses a full distance matrix for each class while the latter involves a single matrix for all the classes. The results published in [26] suggest that MPCKM-S-D requires a high number of constraints to learn the metric. This is particularly the case for full matrices: indeed using diagonal matrices seems to give better results when the number of constraints is low. Conversely to the other algorithms, COP-KMeans requires that all the constraints be satisfied; this explains why this method outperforms the others (except CECM-mah) for a high number of constraints. The drawback is that the constraints may not be consistent with the initial partition and therefore the optimisation problem may be infeasible [1].

[1]The code is available at http://www.cs.utexas.edu/users/ml/risc/

C. Image Segmentation

The interest of CECM will now be illustrated using an example in image segmentation. An image of a plane is given in Figure 3. The aim was to isolate the plane from the rest of the image.

In a first experiment we used ECM. We consider that there is no outlier in the image, so we force the mass on the empty set to be as small as possible by setting ρ^2 to a high value. So starting from the gray levels of the pixels (rescaled between 0 and 1), ECM, with $c = 2$, $\alpha = 1$, and $\rho^2 = 10$ and V initialized with FCM, finds a hard credal partition represented in Figure 3. It may be seen that the ECM fails to isolate properly the plane. In a second experiment, we introduced constraints on the partition as shown in Figure 4. The pixels belonging to the region delimited by hand are mutually linked by a must-link constraint. CECM with an adaptive metric was run using FCM for the initialization of V and with the following parameters: $c = 2$, $\alpha = 1$, $\gamma = 1$, $\rho^2 = 10$. The credal partition is presented in the right part of Figure 4. 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 plane. Note that the remaining pixels allocated to Ω are lying at the boundary between the plane and the sky. As a matter of comparison, are also given in Figure 5 the hard partitions obtained from the pignistic probabilities computed from the results of ECM and CECM.

Fig. 3. (Left) Original image; (Right) Credal partition obtained with ECM (white area: ω_1, grey area: ω_2, black area: Ω).

Fig. 4. (Left) Must-link constraints; (Right) Credal partition obtained with CECM (white area: ω_1, grey area: ω_2, black area: Ω).

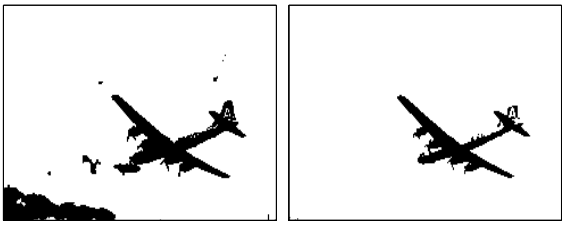

Fig. 5. (Left) Hard partition computed from ECM; (Right) Hard partition computed from CECM.

VI. 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 proved that these two extensions make it possible to guide the algorithm towards desired solutions. Moreover they showed that our algorithm gives good results compared with other constraint-based methods since the former requires less constraints to give satisfying solutions.

References

[1] 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*. Society for Industrial Mathematics, 2005, p. 138.

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

[3] D. Gondek and T. Hofmann, "Non-redundant data clustering," *Knowledge and Information Systems*, vol. 12, no. 1, pp. 1–24, 2007.

[4] 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*, 2001, pp. 577–584.

[5] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," in *Advances in Neural Information Processing Systems 15*. MIT Press, 2003, pp. 505–512.

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

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

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

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

[10] N. Grira, M. Crucianu, and N. Boujemaa, "Active semi-supervised fuzzy clustering," *Pattern Recognition*, vol. 41, no. 5, pp. 1851–1861, 2008.

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

[12] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, 1993.

[13] S. Sen and R. Davé, "Clustering of relational data containing noise and outliers," in *Fuzzy Systems Proceedings*, vol. 2, 1998, pp. 98–110.

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

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

[16] ——, "RECM: Relational evidential c-means algorithm," *Pattern Recognition Letters*, vol. 30, pp. 1015–1026, 2009.

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

[18] R. Davé, "Clustering relational data containing noise and outliers," *Pattern Recognition Letters*, vol. 12, pp. 657–664, 1991.

[19] D. Gustafson and W. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, vol. 17, 1978, pp. 761–765.

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

[21] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, pp. 191–234, 1994.

[22] P. Smets, "The transferable belief model for quantified belief representation," in *Hanbook of Defeasible Reasoning and Uncertainty Management Systems*, vol. 1. Kluwer Academic Publishers, 1998, pp. 267–301.

[23] R. Yager, "On the normalization of fuzzy belief structures," *International Journal of Approximate Reasoning*, vol. 14, no. 2-3, pp. 127–153, 1996.

[24] Y. Ye and E. Tse, "An extension of karmarkar's projective algorithm for convex quadratic programming," *Mathematical Programming*, vol. 44, no. 1, pp. 157–179, 1989.

[25] 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.

[26] M. Bilenko, S. Basu, and R. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proceedings of the twenty-first international conference on Machine learning*. ACM New York, NY, USA, 2004.

[27] I. Davidson and S. Ravi, "Intractability and clustering with constraints," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, p. 208.