ECM: An evidential version of the fuzzy *c*-means algorithm

Marie-Hélène Masson, T. Denœux

PII: S0031-3203(07)00406-2
DOI: doi:10.1016/j.patcog.2007.08.014
Reference: PR 3020

To appear in: *Pattern Recognition*

Received date: 1 February 2007
Revised date: 24 July 2007
Accepted date: 29 August 2007

Cite this article as: Marie-Hélène Masson and T. Denœux, ECM: An evidential version of the fuzzy *c*-means algorithm, *Pattern Recognition* (2007), doi:10.1016/j.patcog.2007.08.014

ECM: An evidential version of the fuzzy c-means algorithm

Marie-Hélène Masson * T. Denœux

UMR CNRS 6599 Heudiasyc, Université de Technologie de Compiègne BP 20529 - F-60205 Compiègne cedex - France

Abstract

A new clustering method for object data, called ECM (Evidential c-means) is introduced, in the theoretical framework of belief functions. It is based on the concept of credal partition, extending those of hard, fuzzy and possibilistic ones. To derive such a structure, a suitable objective function is minimized using a FCM-like algorithm. A validity index allowing the determination of the proper number of clusters is also proposed. Experiments with synthetic and real data sets show that the proposed algorithm can be considered as a promising tool in the field of exploratory statistics.

Key words: Clustering, unsupervised learning, Dempster-Shafer theory, evidence theory, belief functions, cluster validity, robustness

1 Introduction

Cluster analysis is an exploratory data analysis tool which aims at grouping a set of n objects into c clusters $\omega_1, ..., \omega_c$ whose members are similar in some way. To measure their similarity, the objects are either described by *object* data or *relational* data. Object data give an explicit description of the objects using p numeric attributes. Relational data arise from direct pairwise measurement of similarities or dissimilarities between the objects. A wide variety of methods

* Email: mmasson@hds.utc.fr. Fax: (33) 3 44 23 44 77. Tel: (33) 3 44 23 49 28

for clustering object and relational data have been developed. They can be broadly classified into two main families : *hierarchical* and *hard or fuzzy partitioning* methods. Hierarchical (divisive or agglomerative) methods provide a description of the data in the form of a sequence of nested clusters. Using hard partitioning methods, objects are grouped in an exclusive way, so that if a certain object belongs to a cluster then it cannot be included in another cluster. On the contrary, with fuzzy partitioning, each object may belong to two or more clusters with different degrees of membership. The most popular fuzzy partitioning method is Bezdek's Fuzzy C-means (FCM) algorithm [4] for object data (an extension of the fuzzy ISODATA algorithm proposed by Dunn [16]) and its relational counterpart, the so-called relational fuzzy c-means (RFCM) [18] .

By minimizing a suitable objective function, these algorithms compute a fuzzy partition matrix (also called a probabilistic fuzzy partition), i.e. a matrix $U = (u_{ik})$ of size $n \times c$ such that

$$\sum_{k=1}^{c} u_{ik} = 1 \quad \forall\, i \in \{1, \ldots, n\} \tag{1}$$

and

$$\sum_{i=1}^{n} u_{ik} > 0 \quad \forall\, k \in \{1, \ldots, c\}\ . \tag{2}$$

Each number $u_{ik} \in [0, 1]$ is interpreted as a *degree of membership* of object i to cluster k. Several authors, having observed counterintuitive results and a poor robustness against noise and outliers, have proposed to relax the normalization

constraint defined by equation (1). The resulting partition is referred to as a *possibilistic partition*. For instance, Krishnapuram and Keller introduced the possibilistic (PCM) clustering algorithm [21] by modifiying the objective function to be minimized. The membership u_{ik} obtained by PCM is interpreted as a *typicality* degree or a *possibility* degree (in the sense of possibility theory [15]) that object i belongs to cluster k. Using a different approach, Davé [9] proposed to add a *noise* cluster, grouping objects badly represented by the clusters. Initially developed for object data, these algorithms have also been extended to handle relational data [19,27,11].

Recently, a new concept of partition, the *credal* partition, based on the belief functions theory, has been introduced in [13,12]. A credal partition extends the existing concepts of hard, fuzzy (probabilistic) and possibilistic partition by allocating, for each object, a "mass of belief", not only to single clusters, but also to any subsets of $\Omega = \{\omega_1, ..., \omega_c\}$. Experiments have shown that this additional flexibility allows to gain a deeper insight in the data and to improve robustness with respect to outliers. An algorithm to derive the partition from relational data, called EVCLUS (EVidential CLUStering), has been developed. In this paper, we address the problem of computing a credal partition from object data and we propose a new algorithm, called ECM (Evidential c-Means), inspired from FCM and from Davé's Noise-Clustering algorithm.

The rest of the paper is organized as follows. Section 2 recalls the necessary background about belief functions and the main fuzzy partitioning algorithms

from which ECM is derived. Section 3 recalls the definition of a credal partition and explains how to compute such a partition from data. The interpretation of a credal partition and the determination of the number of clusters are discussed and illustrated using synthetic and real data sets in Section 4. An analysis of the complexity of the method and some guideline for choosing the parameters of the method are also presented. Section 5 presents an application of the method to the segmentation of medical images. Finally, Section 6 concludes the paper.

2 Background

2.1 Belief functions

The Dempster-Shafer theory of evidence (or belief functions theory), like probability or possibility theories, is a theoretical framework for reasoning with partial and unreliable information. It encompasses different models of reasoning under uncertainty including Smets's Transferable Belief Model [31]. In this section, only the main concepts of this theory are recalled. A more complete description can be found in Shafer's book [28].

Let us consider a variable ω taking values in a finite set Ω called the frame of discernment. Partial knowledge regarding the actual value taken by ω can be represented by a *basic belief assignment* (bba) [28,30], defined as a function m

from 2^Ω to $[0, 1]$, verifying:

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

The subsets A of Ω such that $m(A) > 0$ are the *focal sets* of m. Each focal set A is a set of possible values for ω, and the number $m(A)$ can be interpreted as a fraction of a unit mass of belief, which is allocated to A on the basis of a given evidential corpus. Complete ignorance corresponds to $m(\Omega) = 1$, whereas perfect knowledge of the value of ω is represented by the allocation of the whole mass of belief to a unique singleton of Ω (m is then called a *certain* bba). When all focal sets of m are singletons, m is equivalent to a probability function, and is called a *Bayesian* bba.

A bba m such that $m(\emptyset) = 0$ is said to be normal. This condition was originally imposed by Shafer [28], but it may be relaxed if one accepts the *open-world assumption* stating that the set Ω might not be complete, and ω might take its value outside Ω [29]. The quantity $m(\emptyset)$ is then interpreted as a mass of belief given to the hypothesis that ω might not lie in Ω.

A bba m can be equivalently represented by a plausibility function $\mathrm{pl} : 2^\Omega \mapsto [0, 1]$, defined as

$$\mathrm{pl}(A) \triangleq \sum_{B \cap A \neq \emptyset} m(B) \quad \forall A \subseteq \Omega \; . \tag{4}$$

The plausibility $\mathrm{pl}(A)$ represents *potential* amount of support given to A. It is important to note that pl boils down to a probability measure when m is a Bayesian bba and to a possibility measure when the focal elements are nested.

Probability and possibility measures are thus recovered as special cases of belief functions.

If two bbas m_1 and m_2 representing distinct items of evidence concerning the value of ω are available, the standard way of combining them is through the conjunctive sum operation $\textcircled{\cap}$ [29] defined as:

$$(m_1 \textcircled{\cap} m_2)(A) \triangleq \sum_{B \cap C = A} m_1(B) m_2(C) \ , \quad \forall A \subseteq \Omega. \tag{5}$$

If necessary, the normality condition $m(\emptyset) = 0$ may be recovered by dividing each mass $(m_1 \textcircled{\cap} m_2)(A)$ by $1 - K$ with $K = (m_1 \textcircled{\cap} m_2)(\emptyset)$. The resulting operation is noted $\oplus$ and is called Dempster's rule of combination [28]:

$$(m_1 \oplus m_2)(A) \triangleq \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \subseteq \Omega, A \neq \emptyset \ . \tag{6}$$

The available evidence being modeled in the form of a basic belief assignment, it is often desirable or necessary to make a decision regarding the selection of one single hypothesis in Ω. In this case, a first solution consists in choosing the singleton in Ω with the highest plausibility [8]. Alternatively, Smets [30] has proposed and justified the use of a probability function. He has shown that the only transformation of a belief function into a probability function satisfying elementary rationality requirements is the pignistic transformation, in which each mass of belief $m(A)$ is equally distributed among the elements of A [32]. This leads to the concept of pignistic probability, BetP, defined, for

a normal bba, by:

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

where $|A|$ denotes the cardinality of $A \subseteq \Omega$. If the bba is subnormal ($m(\emptyset) \neq 0$), then a preliminary normalization step has to be performed. Two methods may apply: Dempster's normalization that consists in dividing all the masses given to nonempty sets by $m(\emptyset)$, and Yager's normalization in which the mass $m(\emptyset)$ is transferred to $m(\Omega)$ [34].

2.2 Fuzzy and possibilistic c-means

Let $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ be a collection of vectors in $\mathbb{R}^p$ describing the n objects. Let c ($2 \leq c < n$) be the desired number of classes. Each cluster is represented by a prototype or a center $\mathbf{v}_k \in \mathbb{R}^p$. Let V denotes a matrix of size $(c \times p)$ composed of the coordinates of the cluster centers such that V_{kq} is the qth component of the cluster center $\mathbf{v}_k$. FCM looks for a partition matrix $U = (u_{ik})$ of size $(n \times c)$ and for the matrix V by minimizing the following objective function:

$$J_{\mathrm{FCM}}(U, V) \triangleq \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{\beta} d_{ij}^2 \quad , \tag{8}$$

subject to the constraints (1) and (2). In the objective function (8), $\beta > 1$ is a weighting exponent that controls the fuzziness of the partition and d_{ij} denotes the Euclidean distance between $\mathbf{x}_i$ and the cluster center $\mathbf{v}_j$. The objective function is minimized using an iterative algorithm, which alternatively optimizes the cluster centers and the membership degrees. The update formulas

are obtained by introducing a Lagrange multiplier with respect to constraint (1) and setting the partial derivatives of the Lagrangian with respect to the parameters to zero [4]. They are given by:

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

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

The algorithm starts from an initial guess for either the partitioning matrix or the cluster centers and iterates until convergence. Convergence is guaranteed but it may lead a local minimum [4].

As it is pointed out, for instance, in [21] or [14], the probabilistic constraint (1) is responsible for undesirable effects, among which the inability to detect noisy data or outliers. In fact, from (10), it can be easily seen that the membership of an object i to cluster j depends not only on d_{ij} but also on its distances to all other clusters. In the possibilistic version of the c-means algorithm introduced by Krishapuram and Keller [21], the probabilistic constraint is dropped and, to avoid the trivial solution $u_{ij} = 0$ for all i and j, a penalty term is added to the objective function:

$$J_{\text{PCM}}(U, V) \triangleq \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{\beta} d_{ij}^2 + \sum_{j=1}^{c} \eta_j \sum_{i=1}^{n} (1 - u_{ij})^{\beta} \quad , \tag{11}$$

where the η_i are fixed, user-specified positive weights, balancing the opposite effects of the two terms in J_{PCM}. An alternating optimization procedure, similar to FCM, can also be derived from the usual first order necessary conditions

for U and V. The update equation for the membership degree is given by

$$u_{ij} = \frac{1}{1 + \left(d_{ij}^2/\eta_i\right)^{1/(\beta-1)}}, \tag{12}$$

whereas the update formula for the centers remains the same than in FCM. Equation (12) shows that the membership degree of an object i to a cluster j depends only on its distance to this cluster, reflecting thereby a *typicality* degree instead of a *relative* membership. This property enables the algorithm to detect atypical data. Several authors, however, have underlined the fact that the independent treatment of the clusters may lead sometimes to unsatisfactory results [2,22]. In fact, PCM exhibits a tendency to produce clusters closer to each other than FCM does. For example, a group of objects can be represented by two or more clusters in the possibilistic model, while some other objects exhibiting a cluster structure are not covered by clusters in the model.

An alternative approach has been proposed by Davé [9] with 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 given by:

$$u_{i*} = 1 - \sum_{k=1}^{c} u_{ik}, \tag{13}$$

relaxing implicitly the probabilistic constraint for the c real clusters. The ob-

jective function to be minimized can be written as:

$$J_{\text{NC}}(U,V) \triangleq \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{\beta} d_{ij}^2 + \sum_{i=n}^{c} \delta^2 \left(1 - \sum_{j=1}^{c} u_{ij}\right)^{\beta}. \tag{14}$$

The algorithm is similar to FCM and PCM with the following updating equation for the membership degrees:

$$u_{ij} = \frac{d_{ij}^{-2/(\beta-1)}}{\sum_{k=1}^{c} d_{ik}^{-2/(\beta-1)} + \delta^{-2/(\beta-1)}}. \tag{15}$$

Note that, as suggested by Davé, the squared value of parameter δ may be fixed as a proportion of the mean of the squared distances between points and cluster centers, after a first run of the algorithm without outlier rejection:

$$\delta^2 = \lambda \frac{1}{c \cdot n} \left(\sum_{i=1}^{n}\sum_{j=1}^{c} d_{ij}^2\right), \tag{16}$$

where λ is a user-defined parameter determining the proportion.

These three algorithms have inspired the clustering method presented in the next section. The proposed approach is developed within the framework of belief functions theory and is based on the concept of credal partition.

3 Evidential c-means

3.1 Credal partition

In [13,12,24], it was proposed to represent partial knowledge regarding the class membership of an object i by a bba m_i on the set $\Omega = \{\omega_1, ..., \omega_c\}$. This representation makes it possible to model all situations ranging from complete

ignorance to full certainty concerning the class of i.

Example. Let us consider a collection of $n = 5$ objects and $c = 3$ classes. Bbas for each object are given in Table 1. They illustrate various situations: the class of object 2 is known with certainty, whereas the class of object 5 is completely unknown; the cases of objects 3 and 4 correspond to situations of partial knowledge (m_4 is Bayesian); finally, the mass $m_1(\emptyset) = 1$ indicates a strong evidence that the class of object 1 does not lie in Ω.

INSERT TABLE I

A credal partition is defined as the n-tuple $M = (m_1, \ldots, m_n)$. It can 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, as defined by Bezdek [5];
- 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.

Note that a credal partition can be converted into a fuzzy or a possibilistic one, as will be shown in Section 4.

A credal partition $M = (m_1, \ldots, m_n)$ is of size c if:

- each bba m_i, $i = 1, \ldots, n$ is defined on a frame Ω of c elements, and
- each class has a strictly positive degree of plausibility for at least one object, i.e., for all $\omega \in \Omega$, we have $\mathrm{pl}_i(\{\omega\}) > 0$ for some $i \in \{1, \ldots, n\}$, pl_i being the plausibility function associated to m_i. Note that this condition is the equivalent of (2) in the definition of a fuzzy c-partition.

3.2 Deriving a credal partition from object data: objective function

Deriving a credal partition from object data implies determining, for each object i, the quantities $m_{ij} = m_i(A_j)$ ($A_j \neq \emptyset, A_j \subseteq \Omega$) in such a way that m_{ij} is low (resp. high) when the distance d_{ij} between i and the focal set A_j is high (resp. low). The distance between an object and any non empty subset of Ω has thus to be defined. Like in fuzzy clustering, we assume that each class ω_k is represented by a center $\mathbf{v}_k \in \mathbb{R}^p$. We propose to associate to each subset A_j of Ω the barycenter $\bar{\mathbf{v}}_j$ of the centers associated to the classes composing A_j. More precisely, introducing the notation

$$s_{kj} = \begin{cases} 1 & \text{if } \omega_k \in A_j \\ 0 & \text{else} \end{cases}, \tag{17}$$

we compute the barycenter $\bar{\mathbf{v}}_j$ associated to A_j by:

$$\bar{\mathbf{v}}_j = \frac{1}{c_j} \sum_{k=1}^{c} s_{kj} \mathbf{v}_k, \tag{18}$$

where $c_j = |A_j|$ denotes the cardinal of A_j. The distance d_{ij} is then defined by:

$$d_{ij}^2 \triangleq ||\mathbf{x}_i - \bar{\mathbf{v}}_j||^2. \tag{19}$$

According to the interpretation of a credal partition explained in the previous section, a separate treatment of the empty set is proposed. This particular focal element is in fact assimilated to a noise cluster, allowing to detect atypical data. Thus, we introduce in the objective function an additional term similar to that in (14) depending on a fixed distance δ between all objects and the empty set.

Finally, we propose to look for the credal partition $M = (m_1, \ldots, m_n) \in \mathbb{R}^{n \times 2^c}$ and the matrix V of size $(c \times p)$ of cluster centers minimizing the following objective function:

$$J_{\text{ECM}}(M, V) \triangleq \sum_{i=1}^{n} \sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} c_j^{\alpha} m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta}, \tag{20}$$

subject to

$$\sum_{\{j/A_j \subseteq \Omega, A_j \neq \emptyset\}} m_{ij} + m_{i\emptyset} = 1 \quad \forall i = 1, n, \tag{21}$$

where $m_{i\emptyset}$ denotes $m_i(\emptyset)$. The criterion J_{ECM} is similar to J_{NC} except that an additional weighting coefficient (c_j^{α}) is introduced: it aims at penalizing the subsets in Ω of high cardinality, the exponent α allowing to control the degree of penalization. Parameters β and δ have the same meaning as in Davé's method. What fundamentally differentiates the two methods is that a credal partition has more degrees of freedom than a fuzzy one. It may thus allow a better modeling and more detailed description of complex data.

Remark 1 A previous attempt to enrich the concept of fuzzy partition was proposed under the name of fuzzy (c+2)-means by Ménard et al [25] who in-

troduced the concept of "ambiguity clusters". Although some ideas are shared by the two methods, the approaches are radically different: the fuzzy (c+2)-means is not formalized within the framework of belief functions, and the geometrical model as well as the optimization process are distinct.

REMARK 2 As indicated in the introduction, an algorithm called EVCLUS [12] was previously developed to derive a credal partition from relational data. Although founded on the same general model of partitioning, EVCLUS and ECM are very different. EVCLUS is applicable to both metric and non metric dissimilarity data and does not use any explicit geometrical model of the data. It only postulates that the more similar two objects, the more plausible it is that they belong to the same cluster. A credal partition is determined in such a way that this condition is, at least approximately, realized. The optimization procedure is based on gradient descent of a stress function and is somewhat related to Multidimensional Scaling (MDS) methods. In contrast, ECM is in line with FCM, PCM and NC: each class is represented by a prototype and the similarity between an object and a cluster is measured using by the Euclidean metric. As will be seen in the next section, ECM uses an alternate optimization procedure to find a credal partition minimizing criterion (20). This makes ECM compuationally much more efficient than EVCLUS when applied to object data.

3.3 Optimization

To minimize J_{ECM}, an alternate optimization scheme can be designed as in the FCM and NC algorithms. First, we consider that V is fixed. To solve the constrained minimization problem with respect to M, we introduce n Lagrange multipliers λ_i and write the Lagrangian:

$$\mathcal{L}(M, \lambda_1, ..., \lambda_n) = J_{\text{ECM}}(M, V) - \sum_{i=1}^{n} \lambda_i \left(\sum_{\{j/A_j \subseteq \Omega, A_j \neq \emptyset\}} m_{ij} + m_{i\emptyset} - 1 \right). \quad (22)$$

By differentiating the Lagrangian with respect to the m_{ij}, $m_{i\emptyset}$ and λ_i and setting the derivatives to zero, we obtain:

$$\frac{\partial \mathcal{L}}{\partial m_{ij}} = \beta c_j^{\alpha} m_{ij}^{\beta-1} d_{ij}^2 - \lambda_i = 0, \quad (23)$$

$$\frac{\partial \mathcal{L}}{\partial m_{i\emptyset}} = \beta \delta^2 m_{i\emptyset}^{\beta-1} - \lambda_i = 0, \quad (24)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = \sum_{\{j/A_j \subseteq \Omega, A_j \neq \emptyset\}} m_{ij} + m_{i\emptyset} - 1 = 0. \quad (25)$$

We thus have from (23)

$$m_{ij} = \left(\frac{\lambda_i}{\beta} \right)^{1/(\beta-1)} \left(\frac{1}{c_j^{\alpha} d_{ij}^2} \right)^{1/(\beta-1)}, \quad (26)$$

and from (24)

$$m_{i\emptyset} = \left(\frac{\lambda_i}{\beta} \right)^{1/(\beta-1)} \left(\frac{1}{\delta^2} \right)^{1/(\beta-1)}. \quad (27)$$

Using (25), (26) and (27),

$$\left(\frac{\lambda_i}{\beta} \right)^{1/(\beta-1)} = \left(\sum_j \frac{1}{c_j^{\alpha/(\beta-1)}} \frac{1}{d_{ij}^{2/(\beta-1)}} + \frac{1}{\delta^{2/(\beta-1)}} \right)^{-1}. \quad (28)$$

Returning in (26), one obtains the necessary condition of optimality for M:

$$m_{ij} = \frac{c_j^{-\alpha/(\beta-1)} d_{ij}^{-2/(\beta-1)}}{\sum_{A_k \neq \emptyset} c_k^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)} + \delta^{-2/(\beta-1)}} \quad \forall i = 1, n \quad \forall j / A_j \subseteq \Omega, A_j \neq \emptyset \tag{29}$$

and

$$m_{i\emptyset} = 1 - \sum_{A_j \neq \emptyset} m_{ij} \quad \forall i = 1, n. \tag{30}$$

Note that these update equations are very similar to those of the NC algorithm (15) except that there are 2^c values m_{ij} to compute instead of $c+1$ fuzzy membership degrees u_{ij}.

Let us now consider that M is fixed. The minimization of J_{ECM} with respect to V is an unconstrained optimization problem. The partial derivatives of J_{ECM} with respect to the the centers are given by:

$$\frac{\partial J_{\text{ECM}}}{\partial \mathbf{v}_l} = \sum_{i=1}^{n} \sum_{A_j \neq \emptyset} c_j^{\alpha} m_{ij}^{\beta} \frac{\partial d_{ij}^2}{\partial \mathbf{v}_l}. \tag{31}$$

$$\frac{\partial d_{ij}^2}{\partial \mathbf{v}_l} = 2(s_{lj})(\mathbf{x}_i - \bar{\mathbf{v}}_j)(-\frac{1}{c_j}). \tag{32}$$

From (31) and (32) we thus have:

$$\frac{\partial J_{\text{ECM}}}{\partial \mathbf{v}_l} = -2 \sum_{i=1}^{n} \sum_{A_j \neq \emptyset} c_j^{\alpha-1} m_{ij}^{\beta} s_{lj} (\mathbf{x}_i - \bar{\mathbf{v}}_j) \tag{33}$$

$$= -2 \sum_{i=1}^{n} \sum_{A_j \neq \emptyset} c_j^{\alpha-1} m_{ij}^{\beta} s_{lj} (\mathbf{x}_i - \frac{1}{c_j} \sum_k s_{kj} \mathbf{v}_k) \quad \forall l = 1, c. \tag{34}$$

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

$$\sum_i \mathbf{x}_i \sum_{A_j \neq \emptyset} c_j^{\alpha-1} m_{ij}^{\beta} s_{lj} = \sum_k \mathbf{v}_k \sum_i \sum_{A_j \neq \emptyset} c_j^{\alpha-2} m_{ij}^{\beta} s_{lj} s_{kj} \quad l = 1, c. \tag{35}$$

Let B be a matrix of size $(c \times p)$ defined by:

$$B_{lq} = \sum_{i=1}^{n} x_{iq} \sum_{A_j \neq \emptyset} c_j^{\alpha-1} m_{ij}^{\beta} s_{lj} = \sum_{i=1}^{n} x_{iq} \sum_{A_j \ni \omega_l} c_j^{\alpha-1} m_{ij}^{\beta} \quad l = 1, c \quad q = 1, p, \quad (36)$$

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

$$H_{lk} = \sum_{i} \sum_{A_j \neq \emptyset} c_j^{\alpha-2} m_{ij}^{\beta} s_{lj} s_{kj} = \sum_{i} \sum_{A_j \supseteq \{\omega_k, \omega_l\}} c_j^{\alpha-2} m_{ij}^{\beta} \quad k, l = 1, c. \quad (37)$$

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

$$HV = B, \quad (38)$$

which can be solved using a standard linear system solver. Table 2 gives an overview of the algorithm.

INSERT TABLE 2

Example (Diamond data set). We illustrate the behavior of ECM using a first simple example inspired from a classical data set [33]. It is composed of twelve objects which are represented in Figure 1. Objects 1 to 11 are part of Windham's data whereas object 12 is an outlier. ECM was run with the following parameters: $\alpha = 1$, $\beta = 2$, $\delta^2 = 20$ and $\epsilon = 10^{-3}$. A 2-credal partition was imposed so that four focal elements have been considered in the optimization process: ω_1, ω_2, Ω and the empty set. The masses are represented in Figure 2, in which $m(\{\omega_1\})$, $m(\{\omega_2\})$, $m(\Omega)$ and $m(\emptyset)$ are plotted against i. It can be seen that the two natural clusters are correctly recovered for points 1 to 11. Point 6 is assigned a high mass to Ω, which reveals that this point

is ambiguous: it could be assigned either to ω_1 or ω_2. Point 12, which can be considered as an outlier, is logically assigned a high mass to the empty set.

INSERT FIGURES 1 and 2

4 Practical issues

4.1 Interpreting a credal partition

A credal partition carries a lot of information about the data. Depending on what is expected from the analysis, different tools may help the user to interpret the results of ECM.

First, the classical clustering structures (possibilistic, fuzzy and hard partitions) can be recovered. A possibilistic partition can be obtained by computing from each bba m_i the plausibilities $\mathrm{pl}_i(\{\omega_k\})$ of the different clusters:

$$\mathrm{pl}_i(\{\omega_k\}) = \sum_{A \cap \{\omega_k\} \neq \emptyset} m_i(A). \tag{39}$$

The value $\mathrm{pl}_i(\{\omega_k\})$ represents the plausibility (or the possibility) that object i belongs to cluster k. In the same way, a probabilistic fuzzy partition may be obtained by calculating the pignistic probability $\mathrm{BetP}_i(\{\omega_k\})$ induced by each bba m_i and interpreting this value as the degree of membership of the object i to cluster k. Finally, a hard partition can be easily obtained by assigning each object to the cluster with highest pignistic probability, or with highest plausibility. Note that points with high masses on the empty set may optionally

be rejected as outliers before hard assignment to the clusters.

Alternatively, and perhaps more interestingly, the concept of credal partition suggests different ways of summarizing the data. One of these ways consists in assigning each object to the *set of clusters* with the highest mass. One then obtains a partition of the n points in at most 2^c groups, where each group corresponds to a set of clusters. This makes it possible to highlight the points that unambiguously belong to one cluster, and the points that lie at the boundary of two or more clusters. More formally, let $X(A_j)$ denote the set of objects for which the mass assigned to A_j is the highest one:

$$X(A_j) = \{i / m_i(A_j) = \max_k m_i(A_k)\}. \tag{40}$$

The $X(A_j)$ for $j = 1, \ldots, 2^c$ define a hard partition of the n objects which will be referred to as a *hard credal partition*.

Following a point of view similar to the one of Lingras [23] using ideas from rough sets theory, it is also possible to characterize each cluster ω_k by two sets: the set of objects which can be classified in ω_k without any ambiguity and the set of objects which could possibly be assigned to ω_k. These two sets will be referred to as the lower and upper approximations of ω_k and denoted, respectively, ω_k^L and ω_k^U. They can be defined using the $X(A_j)$ as:

$$\omega_k^L = X(\{\omega_k\}), \tag{41}$$

and

$$\omega_k^U = \bigcup_{\{j/\omega_k \in A_j\}} X(A_j). \tag{42}$$

The hard credal partition and the associated lower and upper cluster approximations are thus *qualitative* summaries of the clustering results. Such summaries may be argued to be quite intuitive and easier to interpret than purely numerical results such as fuzzy partitions, while being much richer than classical hard partitions. As such, we believe that they are very useful tools for displaying and interpreting the results of ECM.

Example (Diamond data set). Table 3 shows the plausibilities, the pignistic probabilities (computed with Dempster's normalization) and the hard cluster assignments for the different points. Note that both classes are completely plausible for point 6 ($\text{pl}_6(\{\omega_1\}) = \text{pl}_6(\{\omega_2\}) \approx 1$) whereas none of the classes is plausible for point 12 ($\text{pl}_{12}(\emptyset) = 0.09$ and $\text{pl}_{12}(\emptyset) = 0.15$). To define the hard partition, point 6 was arbitrarily assigned to cluster 1. The last column of Table 3 shows the hard assignments to sets of clusters, defining the hard credal partition. It is defined in that case as:

$$X(\emptyset) = \{12\}, \quad X(\{\omega_1\}) = \{1, 2, 3, 4, 5\},$$

$$X(\{\omega_2\}) = \{7, 8, 9, 10, 11\}, \quad X(\{\omega_1, \omega_2\}) = \{6\}.$$

We thus have in that case

$$\omega_1^L = \{1, 2, 3, 4, 5\}, \quad \omega_1^U = \{1, 2, 3, 4, 5, 6\},$$

$$\omega_2^L = \{7, 8, 9, 10, 11\}, \quad \omega_2^U = \{6, 7, 8, 9, 10, 11\}.$$

INSERT TABLE 3

Example (Four-class data set). To illustrate this interest of the lower and outer cluster approximations in a more realistic situation, let us consider the following synthetic example: four classes of 200 points each are generated from a multivariate t distribution with 5 degrees of freedom and centered respectively on [0;0], [0;4], [4;0] and [4;4]. This data set is represented in Figure 3. The clustering task is to find a four-credal partition with the following parameters: $\alpha = 1$, $\beta = 2$, $\delta^2 = 20$ and $\epsilon = 10^{-3}$. Figure 4 represents the hard credal partition (note that ω_{ij} means $\{\omega_i, \omega_j\}$ and ω_{ijk} means $\{\omega_i, \omega_j, \omega_k\}$). Each subset is represented by its convex hull. The center of gravity $\mathbf{v}_k$ of each cluster is marked by a cross. The points in $X(\emptyset)$ for which the highest mass is given to the empty set are identified by squares. It can be seen that a meaningful partition is recovered and that outliers are correctly detected. Figures 5 and 6 show the lower and upper approximations of each cluster, also represented by their convex hulls.

INSERT FIGURES 3, 4, 5 and 6

4.2 Limiting the complexity

For each object, the ECM algorithm distributes a fraction of the unit mass to each element of 2^Ω. Consequently, the number of parameters to be optimized is exponential in the number of clusters (and linear in the number of objects). For a limited number of classes (say, less than 10), calculations are easily tractable. For example, for the diamond data, each run of the algorithm implemented in Matlab takes about 0.06 seconds on a PC equipped with a Pentium 4 processor.

However, if necessary, it is also possible to reduce the complexity of the method by considering only a subclass of bbas with a limited number of focal sets. For example, we may constrain the focal sets to be either Ω, or to be composed of at most two classes, thereby reducing the complexity from 2^c to c^2. By this way, the number of parameters to be optimized is drastically reduced and an acceptable trade-off between flexibility of the method and computational tractability is achieved.

In Table 4, a comparison between these two versions of ECM (V1: no limitation of the number of focal elements; V2: cardinality of the focal elements at most equal to 2 except for Ω) is provided. This table shows the execution times and the final value of objective function (20) obtained for the four-class data set (800 points). The number of clusters was varied between 2 to 6, and the following settings were used: $\alpha = 2$, $\beta = 2$, $\delta^2 = 20$ and $\epsilon = 10^{-3}$. Mean values

and standard deviations over 100 runs of the algorithm are reported. Results with version V2 are reported only for $c > 3$, as the two versions are identical for $c \leq 3$.

Several remarks can be made. First, one can see that, as expected, J_{ECM} decreases as c grows. In contrast, this monotonic tendency with respect to c is not systematically observed for the mean execution time since it depends also on the adequation between the model and the structure of the data. A significant reduction of computing time is obtained using the constrained version V2, at the price of a slightly higher value of the objective function at convergence, which is obviously due to the lower number of free parameters in the constrained model. Finally, it can be noticed that the results are quite stable, particularly for $c = 4$ which corresponds to the true number of clusters. This suggests that the stability of the partition can be used as a clue for choosing the number of cluster. Another approach to model complexity determination based on a validity index will be presented in the next section.

INSERT TABLE 4

To conclude this section on complexity issues, we may notice that a common strategy for accelerating iterative clustering procedures is to start from a good initial condition obtained using a simpler method. Here, one could think of initializing the centers using FCM instead of random initialization. This strategy, however, did not yield any significant reduction of execution time when

applied to the four-class data.

4.3 Determining the number of clusters

One of the fundamental issue in fuzzy clustering is the choice of a suitable number of clusters. This problem is often referred to as cluster validity. Most of the methods consists in computing a validity index from several partitions obtained with different values of c and looking for a minimum, a maximum or an abrupt change in the criterion. A great number of validity indexes have been proposed for assessing the quality of a fuzzy partition. We review some of them and then propose a new index for assessing the quality of a credal partition.

The first index associated with FCM was the *partition coefficient* defined as [4]:

$$PC(c) \triangleq \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}^2, \tag{43}$$

where $\frac{1}{c} \leq PC(c) \leq 1$. The *partition entropy* [3] was defined by:

$$PE(c) \triangleq -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\mu_{ij}\log_2(\mu_{ij}), \tag{44}$$

where $0 \leq PE(c) \leq \log_2(c)$. The optimal partition (or equivalently the optimal number of clusters c) is obtained by maximizing PC (or minimizing PE) with respect to $c = 2, 3, ..., c_{\max}$. Both PC and PE possess a monotonic tendency with respect to c. To overcome this problem, Davé [10] proposed a modified

PC (MPC) index defined by:

$$MPC(c) \triangleq 1 - \frac{c}{c-1}(1 - PC(c)), \tag{45}$$

where $0 \leq PE(c) \leq 1$. This index has to be maximized.

These indexes have inspired the validity index proposed in the sequel. Intuitively, one feels that if a proper number of classes is chosen, the centers of gravity will cover correctly the clusters and the major part of the mass will be allocated to singletons of 2^Ω. On the contrary, if c is too small or too high, the mass will be distributed to subsets with higher cardinality or to the empty set. In other words, the more specific a credal partition will be, the more accurately it will represent the structure of the data. These remarks lead to the use of one of the entropy measures proposed for belief functions, the *nonspecificity* [26]. Klir and Wierman [20, p. 51] defined the nonspecificity of a subnormal bba m as:

$$N(m) \triangleq \sum_{A \in 2^\Omega \setminus \emptyset} m(A) \log_2 |A| + m(\emptyset) \log_2 |\Omega|, \tag{46}$$

where $0 \leq N(m) \leq \log_2(|\Omega|)$. This measure tends to be small when the mass is assigned to few non empty focal sets with small cardinality. We propose to define the validity index of a credal partition as the average normalized specificity:

$$N^*(c) \triangleq \frac{1}{n \log_2(c)} \sum_{i=1}^{n} \left[\sum_{A \in 2^\Omega \setminus \emptyset} m_i(A) \log_2 |A| + m_i(\emptyset) \log_2(c) \right], \tag{47}$$

where $0 \leq N^*(c) \leq 1$. This index has to be minimized.

Its application is illustrated using three data sets.

Example (Four-classes data set). ECM was run with $\beta = 2$, $\delta^2 = 20$ and $\epsilon = 10^{-3}$. For different values of α (namely 1, 2 and 3), the number of desired clusters was varied from 2 to 6. The results are reported in Figure 7. It can be seen that the minimum is always obtained for $c = 4$ clusters, this minimum being less pronounced as α increases, or, equivalently, as the focal elements of high cardinality are more strongly penalized in the objective function.

INSERT FIGURE 7

Example (Iris data set). This famous data set [1,17] is composed of flowers from the iris species setosa, versicolor, and virginica. From each species there are 50 observations for sepal length, sepal width, petal length, and petal width in cm. Two clusters are known to have a significant overlap so that $c = 2$ or $c = 3$ may be a good choice for these data. The proposed validity index, computed for different α with $\beta = 2$, $\delta^2 = 10$ and $\epsilon = 10^{-3}$, is shown in Figure 8. Depending on α, one can see that a minimum of 2 or 3 is always found.

INSERT FIGURE 8

Example (Soybean data set). This data set on diseases in soybeans, available from the UCI Machine Learning repository, contains 47 data points. Each data point has 35 categorical attributes and is classified into four diseases. One

disease has 17 data, and the three other diseases have 10 data each. The data set is represented using a MDS algorithm in Figure 9. Although presented as a four-class problem, this representation suggests that three clusters could be a reasonable choice for these data. Figure 10 displays the value of $N^*(c)$ obtained for different α with $\beta = 2$, $\delta^2 = 30$ and $\epsilon = 10^{-3}$. A proper minimum is reached for $c = 3$ clusters except for $\alpha = 2.5$, which indicates a cluster number estimate equal to 4.

INSERT FIGURES 9 AND 10

These simple examples suggests that N^* can be a valuable index for credal partitions.

4.4 Some guidelines for choosing the parameters

Before running ECM, one has to set the values of several parameters. As in FCM, PCM or NC, for which it is a usual choice, we used $\beta = 2$ in all experiments. Parameter α allows to control the amount of points assigned to focal elements of high cardinality. The value $\alpha = 2$ can be used as a starting default value but it can be modified according to what is expected from the user: the higher α, the less imprecise will be the resulting partition. The choice of δ is more difficult and is strongly data-dependent. If the number of clusters is fixed, a strategy similar to the one of Davé can be applied, so as to achieve a given rejection rate. For choosing c, we recommend the use of several tools

including the validity index described above, the observation of the variability of the results over several runs as discussed in Section 4.2, but also low dimensional graphical displays of the data provided by, e.g., Principal Component Analysis or MDS.

5 Application to medical image segmentation

The interest of ECM will now be illustrated using an example in medical imaging taken from [6,7]. Two images of a pathological brain were acquired using magnetic resonance imaging using two different sets of acquisition parameters (dual-echo MRI). They are represented in Figure 11. In the first echo, according to the grey levels of the pixels, two main areas may be distinguished: normal brain tissues on the one hand (bright area) and ventricles and cerebrospinal fluid (CSF) (darker areas) on the other hand. The pathology cannot be seen in the first image. In the second echo, the pathology (bright area) is easily seen but CSF and ventricles have almost the same grey levels than the rest of the brain and have ill-defined contours. Several experiments were conducted with these data; they are reported in the following.

INSERT FIGURE 11

We first applied ECM on each image with the following settings: $\alpha = 2$, $\beta = 2$, $\delta^2 = 10$ and $\epsilon = 10^{-3}$ (note that, to avoid manipulating large values of distances between points, the grey level of each pixel was divided by 100 before

processing). The results of these segmentations are presented in Figures 12 and 13 in the form of lower (grey) and upper (grey + dark grey) approximations of the clusters. One can see that each area of interest of the brain (CSF and ventricles, normal tissues, pathology) are well recovered by the lower approximations. The darker areas correspond to doubtful pixels with intermediate grey levels.

INSERT FIGURES 12 and 13

We then considered jointly the two echos and performed a direct classification in three clusters in the two-dimensional space formed by the grey levels of images 1 and 2 . The resulting segmentation (in the form of the lower and upper approximations of the clusters) is shown in Figure 14. Here again, the main areas of interest are well recovered by the lower approximations of the clusters, whereas upper approximations highlight contour or doubtful pixels. Another view of this segmentation, in the form of a hard credal partition in the two-dimensional intensity space, is given in Figure 15.

INSERT FIGURES 14 and 15

The last experiment is intended to illustrate another way of exploiting the great expressive power of a credal partition. As indicated in Section 2.1, an interesting feature of belief function theory is the existence of simple tools for combining different sources of informations. In the segmentation problem considered here, each echo can be considered as a distinct source of information.

It was thus interesting to see if a meaningful three-class partition of the brain could be recovered by combining the individual two-class partitions extracted from each image. Before being combined, the masses have to be expressed on the same frame of discernment.

This problem can be formalized as follows. Let $\Gamma = \{\gamma_1, \gamma_2\}$ denote the frame of discernment of bbas obtained from the first image. Singleton γ_1 corresponds to the ventricles and CSF, γ_2 corresponds to the normal and pathological tissues. In the same way, let $\Theta = \{\theta_1, \theta_2\}$ denote the frame of discernment of bbas associated to the second image with θ_1 corresponding to the pathology, ventricles and CSF and θ_2 to the normal tissues. Finally, let $\Omega = \{\omega_1, \omega_2, \omega_3\}$ be the frame of discernment corresponding to the three areas of interest: singleton ω_1 denotes the pathology, ω_2 denotes the ventricles and CSF, and ω_3 corresponds to the normal brain tissues. Bbas expressed on Γ, Θ, and Ω are respectively denoted m^Γ, m^Θ,and m^Ω.

We can observe that frame Ω is *finer*, i.e., more detailed than the other two: using the terminology introduced by Shafer [28, page 115], it is a *refinement* of Γ and Θ. As a consequence, bbas m^Γ and m^Θ may be expressed on Ω using an operation called a *refining*. It is a very simple process which consists in transferring the masses in the following way. We can see that γ_1 corresponds to ω_2 and γ_2 corresponds to $\{\omega_1, \omega_3\}$. Consequently, $m^\Gamma(\{\gamma_1\})$ may be transferred to $\{\omega_2\}$, $m^\Gamma(\{\gamma_2\})$ to $\{\omega_1, \omega_3\}$, and $m^\Gamma(\Gamma)$ to Ω. The resulting bba m_1^Ω on Ω

is thus defined as

$$m_1^\Omega(\{\omega_2\}) = m^\Gamma(\{\gamma_1\}), \quad m_1^\Omega(\{\omega_1, \omega_3\}) = m^\Gamma(\{\gamma_2\}), \quad m_1^\Omega(\Omega) = m^\Gamma(\Gamma).$$

Similarly, θ_1 corresponds to $\{\omega_1, \omega_2\}$ and θ_2 corresponds to ω_3. Consequently, $m^\Theta(\{\theta_1\})$ can be transferred to $\{\omega_1, \omega_2\}$, $m^\Theta(\{\theta_2\})$ to $\{\omega_3\}$, and $m^\Theta(\Theta)$ to Ω. The resulting bba m_2^Ω on Ω is thus defined as

$$m_2^\Omega(\{\omega_1, \omega_2\}) = m^\Theta(\{\theta_1\}), \quad m_2^\Omega(\{\omega_3\}) = m^\Theta(\{\theta_2\}), \quad m_2^\Omega(\Omega) = m^\Theta(\Theta).$$

Once the bbas have been expressed on the same frame of discernment, they can be combined using Dempster's rule (6) to produce the final mass assignment $m^\Omega = m_1^\Omega \oplus m_2^\Omega$. The result of this combination scheme is presented in Figure 16. The partition obtained is very similar to the previous one, and again all regions of interest are correctly recovered.

INSERT FIGURE 16

This example demonstrates an additional tool offered by the theoretical framework of belief functions for clustering: the possibility to combine in a meaningful way several partitions obtained from different sources. Thanks to this possibility, complex problems in pattern recognition or image analysis may be decomposed into smaller, simpler problems, the solutions of which being fused to derive a global solution. For example, in some cases, some subsets of features are known to be more likely to discriminate some of the classes than others. It is then possible to compute partitions from different subsets

of features (which constitute simpler clustering tasks), and then to combine them. A similar combination approach may also be used for incorporating prior information provided by an expert.

6 Conclusion

A new clustering method based on belief functions theory has been proposed. The concepts of fuzzy, possibilistic and probabilistic partitions have been shown to be recovered as special cases of a more general clustering concept: the credal partition. To derive a credal partition from object data, an efficient algorithm, called ECM, similar to FCM, has been introduced. It is based on a classical alternating minimization scheme, with, in a first step, the determination of the centers of the clusters, and, in a second step, the allocation of the masses to the different subsets of classes. Examples have shown that meaningful partitions of the data could be obtained. A credal partition may be analyzed by computing from it a hard, a possibilistic, or a fuzzy partition as by-products. Alternatively, a hard credal partition can be determined, from which lower and upper approximations of clusters may be computed, providing an intuitive summary of the data. A validity index, based on the notion of nonspecificity of belief functions, has been proposed and illustrated using three data sets. Finally, an application to multimodal image segmentation has been presented. This work offers several perspectives, among which a kernelized version of the algorithm and a new algorithm for relational data.

Acknowledgements

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

References

[1] E. Anderson. The irises of the Gasp peninsula, *Bulletin of the American Iris Society*, 59, 2-5, 1935.

[2] M. Barni, and V. Cappellini, and A. Mecocci. Comments on a possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 4(3), 393-396, 1996.

[3] J. C. Bezdek. Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3, 58-73, 1974.

[4] J. C. Bezdek. *Pattern Recognition with fuzzy objective function algorithms.* Plenum Press, New-York, 1981.

[5] J. C. Bezdek, J. Keller, R. Krishnapuram, and N. R. Pal. *Fuzzy models and algorithms for pattern recognition and image processing.* Kluwer Academic Publishers, Boston, 1999.

[6] I. Bloch. Some aspects of Dempster-Shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account. *Pattern recognition Letters*, 17(8), 905-919, 1996.

[7] I. Bloch. Defining belief functions using mathematical morphology – Application to image fusion under imprecision. *International Journal of Approximate Reasoning*, In press, 2007.

[8] B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3), 314-330, 2006.

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

[10] R.N. Davé. Validating fuzzy partition obtained through *c*-shell clustering. *Pattern Recognition Letters*, 17, 613-623, 1996.

[11] M. De Caceres, and F. Oliva, and X. Font. On relational possibilistic clustering. *Pattern Recognition*, 39, 2010-2024, 2006.

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

[13] T. Denœux, and M.-H. Masson. Clustering of proximity data using belief functions. B. Bouchon-Meunier, L. Foulloy and R. R. Yager, Eds. *Intelligent systems for information processing from representation to application*. 291-302, Elsevier, Amsterdam, 2003.

[14] C. Döring, and M.-J. Lesot, and R. Kruse. Data analysis with fuzzy clustering methods. *Computational and Data Analysis*, 51, 192-214, 2006.

[15] D. Dubois and H. Prade. *Possibility Theory: An approach to computerized processing of uncertainty*. Plenum Press, New-York, 1988.

[16] J.C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters. *Journal of Cybernetics*, 32-57, 1974.

[17] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188, 1936.

[18] R.J. Hathaway, and J.W. Davenport, and J.C. Bezdek. Relational duals of the c-means clustering algorithms. *Pattern Recognition*, 22, 205-212, 1989.

[19] R.J. Hathaway, and J.C. Bezdek, and J.W. Davenport. On relational versions of c-means algorithm. *Pattern Recognition Letters*, 17, 607-612, 1996.

[20] G.J. Klir, and M.J. Wierman. *Uncertainty-based Information. Elements of Generalized Information Theory*. Springer-Verlag, New-York, 1998.

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

[22] R. Krishnapuram, and J. Keller. The possibilistic c-means algorithm: insights and recommendations. *IEEE Transactions on Fuzzy Systems*, 4(3), 385-393, 1996.

[23] P. Lingras. Unsupervised rough set classification using GAs. *Journal of Intelligent Information Systems*, 16, 215-228, 2001.

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

[25] M. Ménard, and C. Demko, and P. Loonis. The fuzzy c+2-means: solving the ambiguity rejection in clustering. *Pattern Recognition*, 33, 1219-1237, 2000.

[26] A. Ramer. Uniqueness of information measure in the theory of evidence. *Fuzzy Sets and Systems*, 24, 183-196, 1987.

[27] S. Sen, and R.N. Davé. Clustering relational data containing noise and outliers. In *FUZZ'IEEE 98*, 1411-1416, 1998.

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

[29] P. Smets. The combination of evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5), 447-458, 1990.

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

[31] P. Smets. The Transferable Belief Model for quantified belief representation. In D.M. Gabbay and P. Smets, Ed. *Hanbook of defeasible reasoning and uncertainty management systems*. volume 1, 267-301, Kluwer Academic Publishers, Dordrecht, 1998.

[32] P. Smets. Decision making in the TBM: the necessity of the pignistic transformation *International Journal of Approximate Reasoning*, 38(2), 133-147, 2005.

[33] M.P. Windham. Numerical classification of proximity data with assignment measure. *Journal of Classification*, 2, 157-172, 1985.

[34] R. R. Yager. On the normalization of fuzzy belief structure. *International Journal of Approximate Reasoning*, 14, 27-153,1996.

List of Figures

13 Image 2. (a) Lower (light grey) and upper (light grey+dark grey) approximations of class 1 (pathology, ventricles and CSF); (b) Lower (light grey) and upper (light grey + dark grey) approximations of class 2 (normal brain tissues). 54

14 Segmentation results obtained using images 1 and 2 simultaneously. (a) pathology; (b) CSF and ventricles; (c) normal brain tissues. The lower approximations of the clusters are represented by light grey areas, the upper approximations by the union of light and dark grey areas. 55

15 Representation of the hard credal partition of the brain data in the two-dimensional intensity space (ω_1=pathology; ω_2=CSF and ventricles; ω_3=normal brain tissues). The focal sets are represented by different grey levels. 56

16 Segmentation of the brain by combination of the partitions obtained from individual images. (a) pathology; (b) CSF and ventricles; (c) normal brain tissues. The lower approximations of the clusters are represented by light grey areas, the upper approximations by the union of light and dark grey areas. 57

A	$m_1(A)$	$m_2(A)$	$m_3(A)$	$m_4(A)$	$m_5(A)$
$\emptyset$	1	0	0	0	0
$\{\omega_1\}$	0	0	0	0.2	0
$\{\omega_2\}$	0	1	0	0.4	0
$\{\omega_1, \omega_2\}$	0	0	0	0	0
$\{\omega_3\}$	0	0	0.2	0.4	0
$\{\omega_1, \omega_3\}$	0	0	0.5	0	0
$\{\omega_2, \omega_3\}$	0	0	0	0	0
Ω	0	0	0.3	0	1

Table 1
Example of credal partition

Input	$\{\mathbf{x}_1, ..., \mathbf{x}_n\}$: n vectors in $\mathbb{R}^p$
Parameters	c: number clusters $1 < c < n$
	$\alpha \geq 0$: weighting exponent for cardinality (default value 2)
	$\beta > 1$: weighting exponent (default value 2)
	$\delta > 0$: distance to the empty set
	ϵ: termination threshold (default value 10^{-3})
Initialization	Choose randomly c initial cluster centers = V_0
Loop	$t \leftarrow 0$
	Repeat
	$t \leftarrow t + 1$
	Compute M_t using (19), (18), (29), (30), and V_{t-1};
	Compute H_t and B_t using (36), (37), and M_t;
	Solve $H_t V_t = B_t$;
	Until $\|\|V_t - V_{t-1}\|\| < \epsilon$

Table 2
Evidential c-means algorithm

i	$\mathrm{pl}_i(\{\omega_1\})$	$\mathrm{pl}_i(\{\omega_2\})$	$\mathrm{BetP}_i(\{\omega_1\})$	$\mathrm{BetP}_i(\{\omega_2\})$	Hard partition	Hard credal partition
1	0.8846	0.0603	0.9534	0.0466	1	1
2	0.8488	0.1195	0.9091	0.0909	1	1
3	0.9973	0.0025	0.9984	0.0016	1	1
4	0.8478	0.1193	0.9091	0.0909	1	1
5	0.8423	0.3803	0.7544	0.2456	1	1
6	0.9994	0.9994	0.5000	0.5000	1	1,2
7	0.4172	0.8388	0.2676	0.7324	2	2
8	0.1291	0.8394	0.0988	0.9012	2	2
9	0.0052	0.9945	0.0034	0.9966	2	2
10	0.1189	0.8534	0.0898	0.9102	2	2
11	0.0563	0.8946	0.0431	0.9569	2	2
12	0.0957	0.1530	0.3628	0.6372	2	$\emptyset$

Table 3
Various partitions obtained using ECM on the Diamond data set: plausibilities (possibilistic partition), pignistic probabilities (probabilistic fuzzy partition), hard partition and hard credal partition.

c	J_{ECM-V1}	J_{ECM-V2}	T_{ECM-V1}	T_{ECM-V2}
2	2548.8 ± 46.2		3.1 ± 1.3	
3	1355.3 ± 10.9		11.8 ± 5.8	
4	720.9 ± 0.00	777.4 ± 0.00	12.2 ± 2.8	7.5 ± 0.02
5	496.2 ± 2.3	581.3 ± 1.8	63.6 ± 31.4	46.1 ± 39.9
6	327.2 ± 3.4	437.5 ± 0.7	185.0 ± 81.3	38.5 ± 14.3

Table 4
Four-classes data set. Values of criterion J_{ECM} and execution times in seconds (V1: no limitation of the focal elements, and V2: with focal elements of size 2). Mean values and standard deviations were computed over 100 runs of the algorithm.

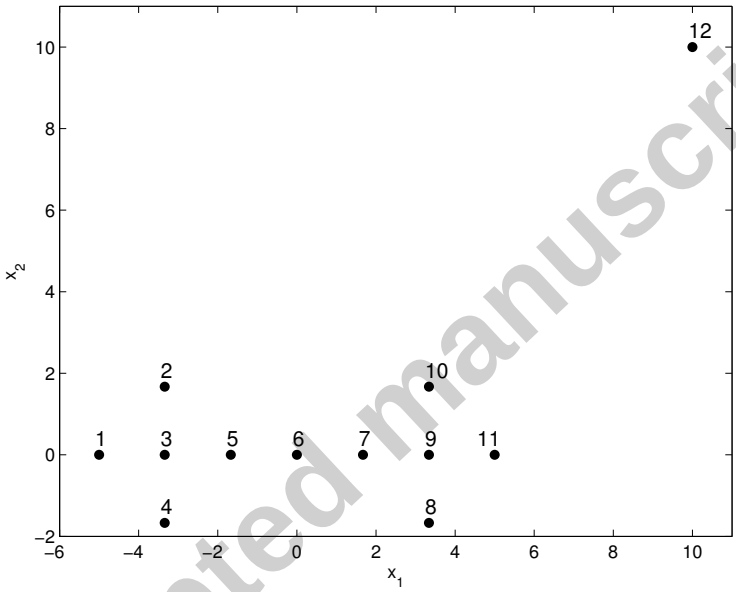


Fig. 1. Diamond data set.

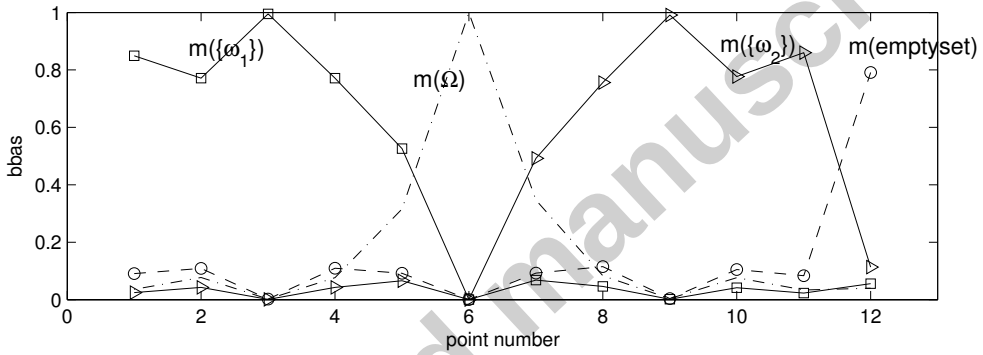


Fig. 2. Credal partition obtained from Diamond data set.

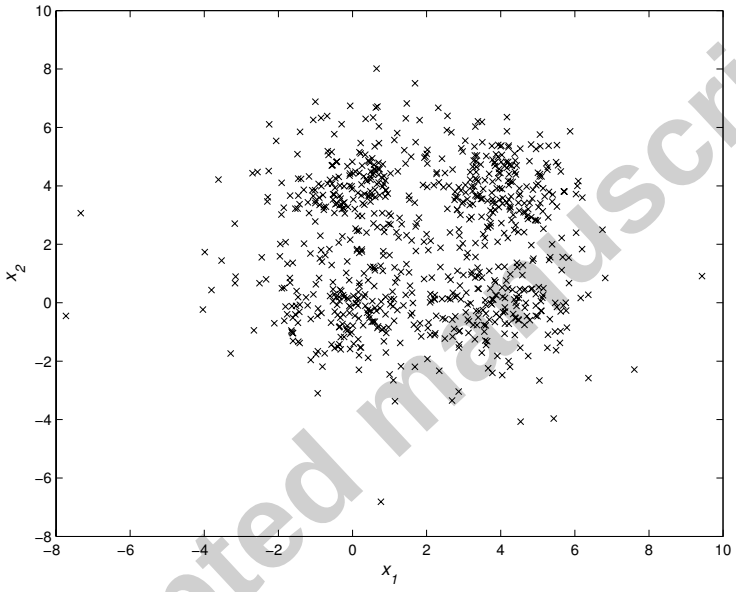


Fig. 3. Four-class data set.

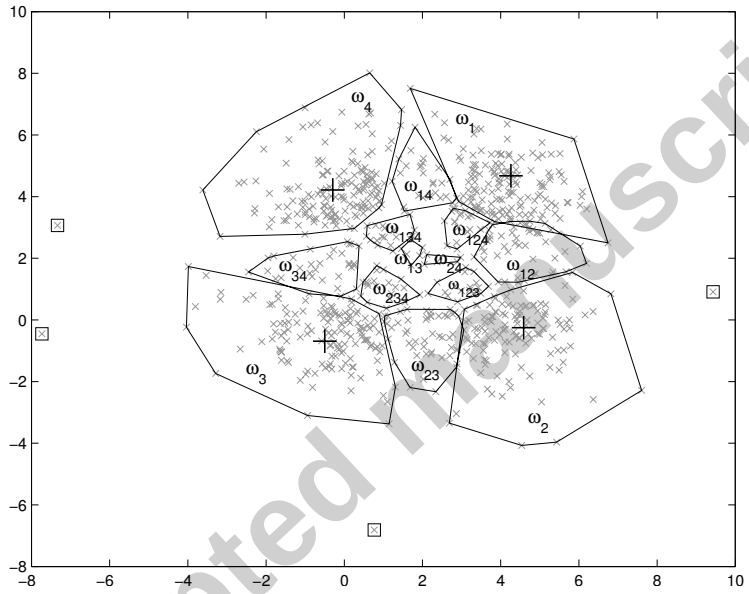


Fig. 4. Four-class data set: hard credal partition computed from m.

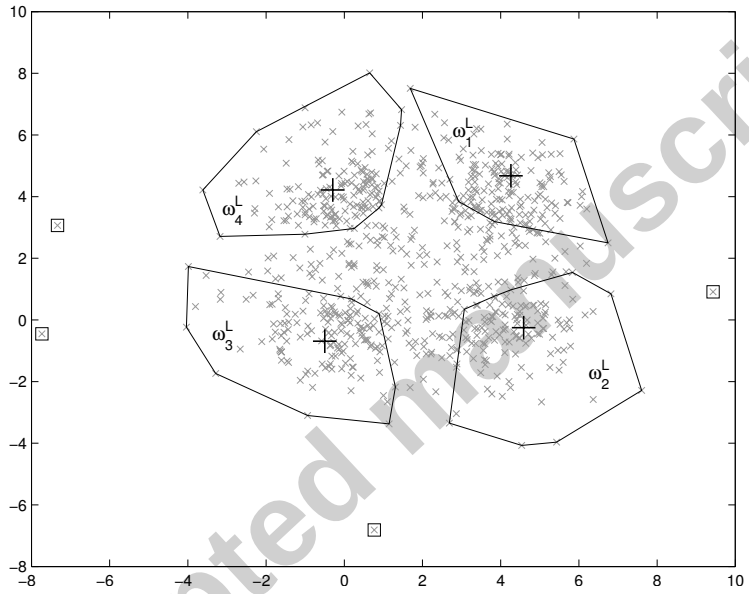


Fig. 5. Four-class data set: lower approximations of the four clusters.

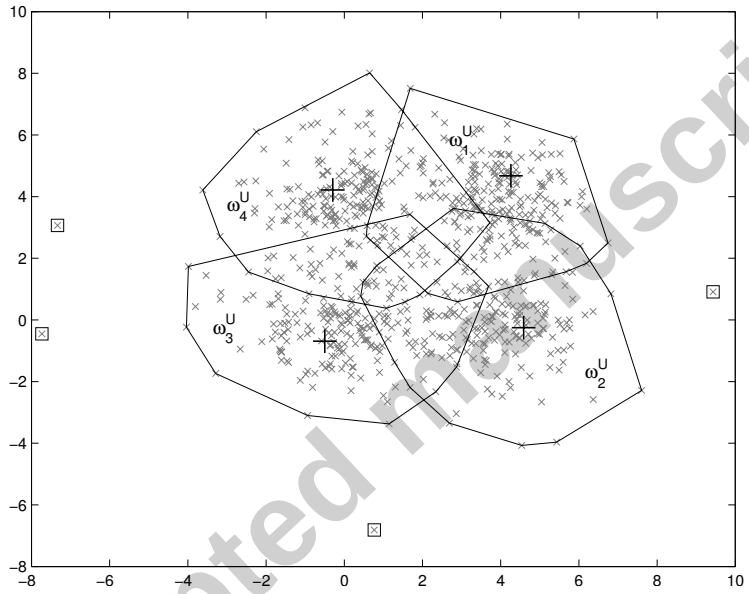


Fig. 6. Four-class data set: upper approximations of the four clusters.

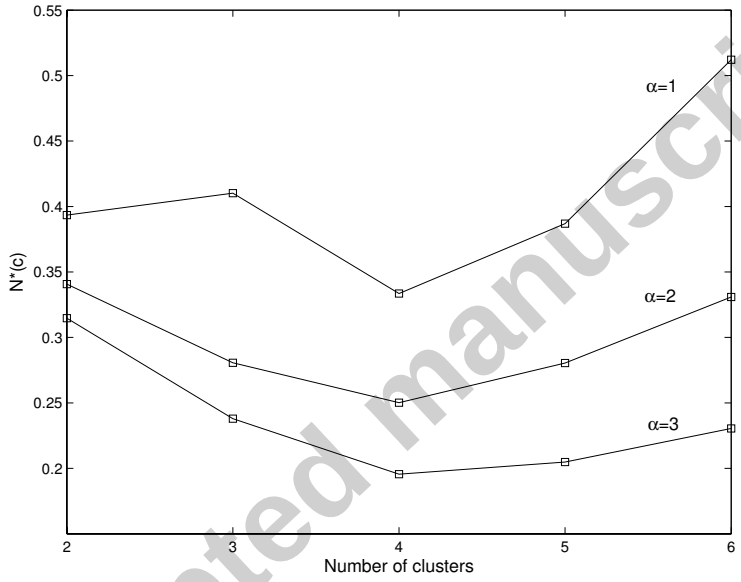


Fig. 7. Four-class data set: validity index.

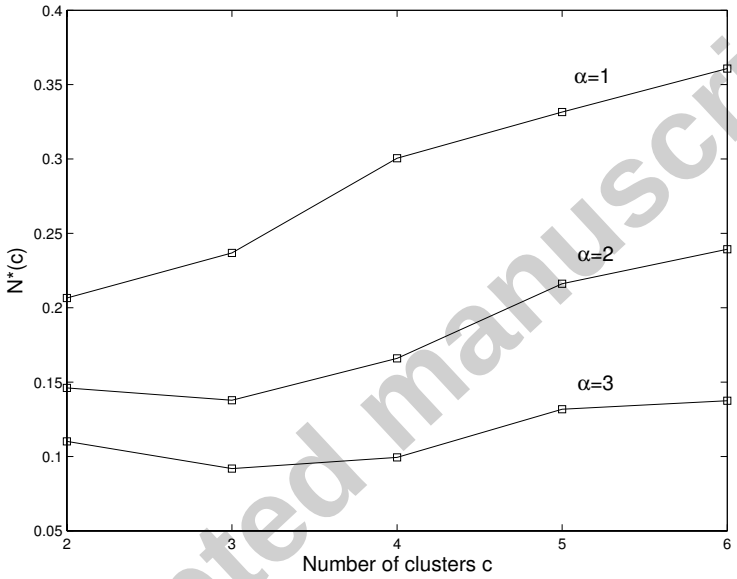


Fig. 8. Iris data set: validity index.

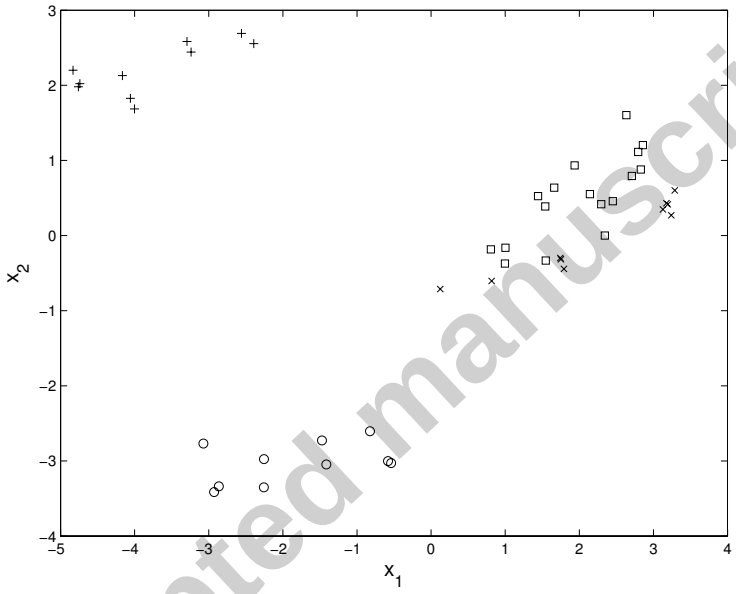


Fig. 9. Soybean data set: MDS representation of the data.

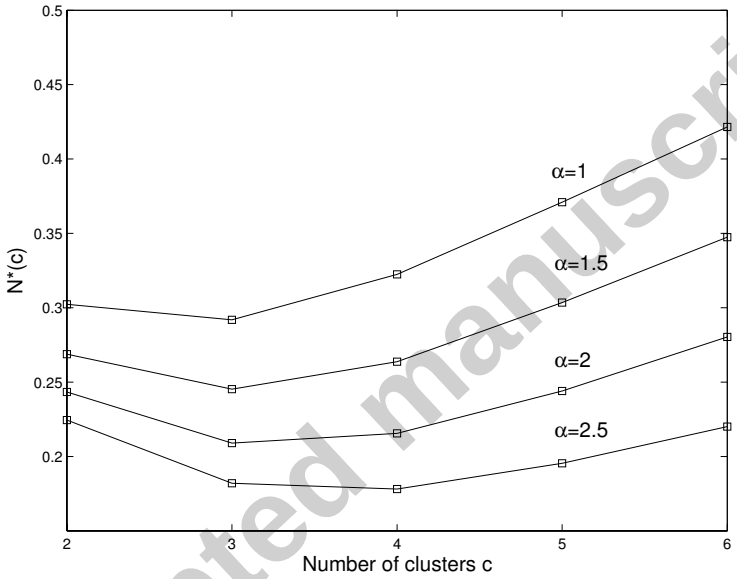
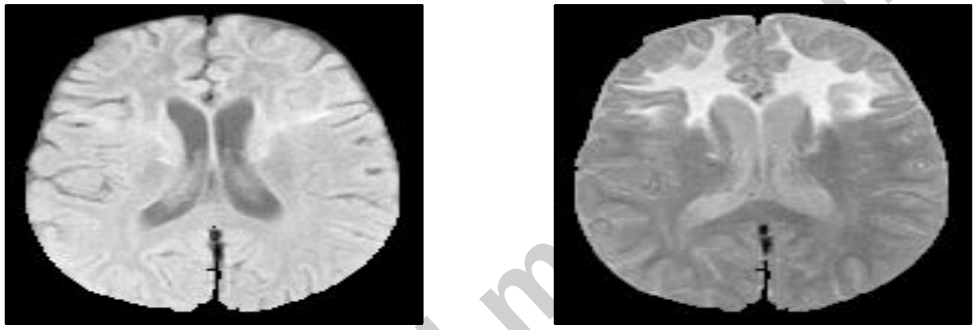


Fig. 10. Soybean data set: validity index.



(a) (b)

Fig. 11. Dual-echo MRI acquisitions.

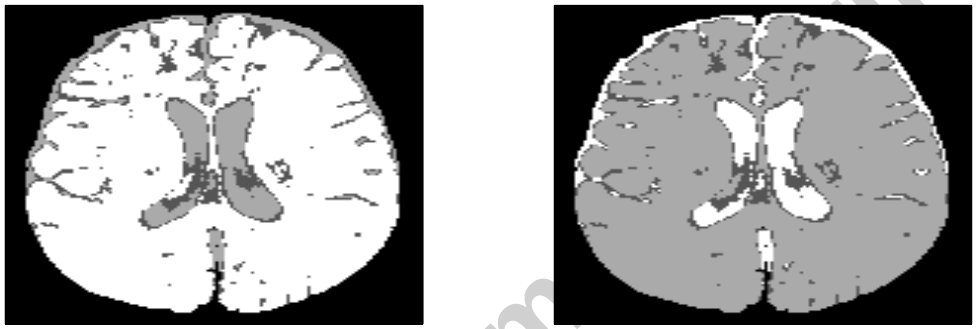


Fig. 12. Image 1. (a) Lower (light grey) and upper (light grey+dark grey) approximations of class 1 (ventricles and CSF); (b) Lower (light grey) and upper (light grey+dark grey) approximations of class 2 (normal and pathological brain tissues).

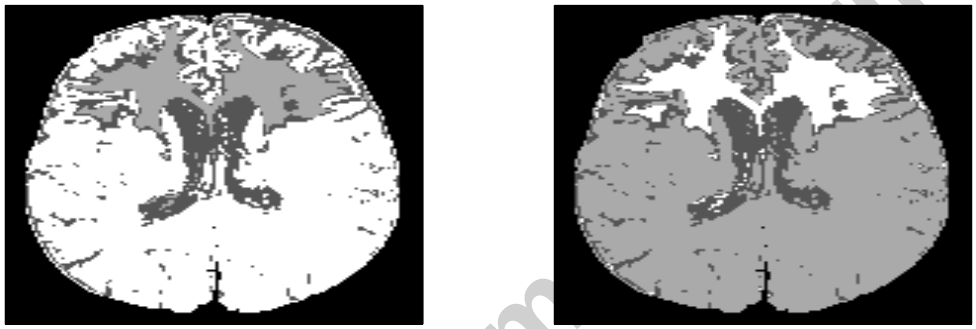


Fig. 13. Image 2. (a) Lower (light grey) and upper (light grey+dark grey) approximations of class 1 (pathology, ventricles and CSF); (b) Lower (light grey) and upper (light grey + dark grey) approximations of class 2 (normal brain tissues).

Fig. 14. Segmentation results obtained using images 1 and 2 simultaneously. (a) pathology; (b) CSF and ventricles; (c) normal brain tissues. The lower approximations of the clusters are represented by light grey areas, the upper approximations by the union of light and dark grey areas.

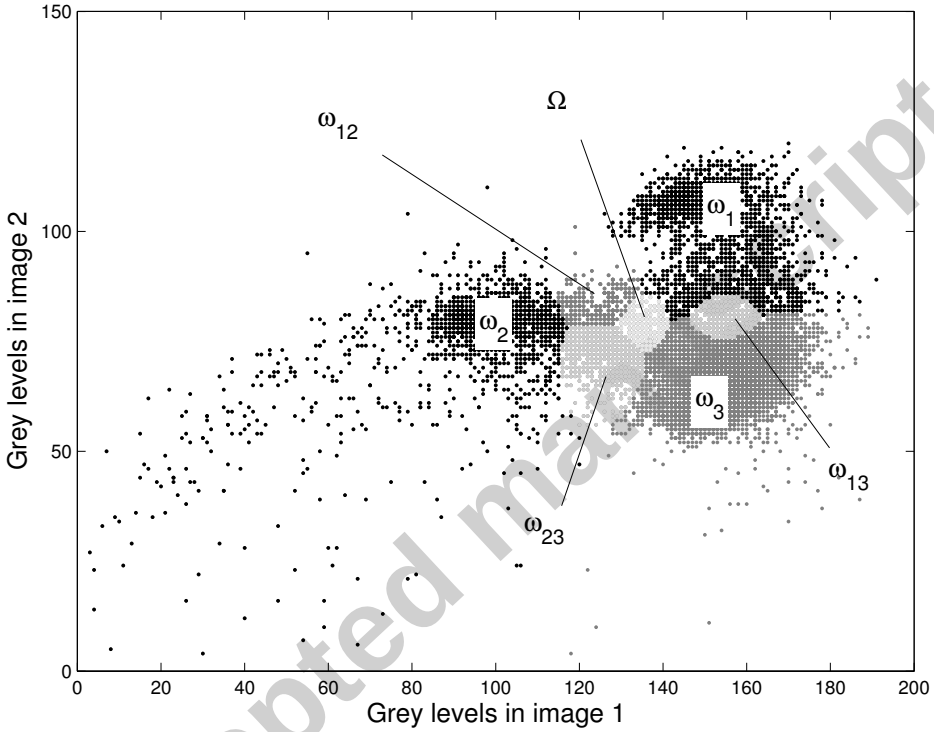


Fig. 15. Representation of the hard credal partition of the brain data in the two-dimensional intensity space (ω_1=pathology; ω_2=CSF and ventricles; ω_3=normal brain tissues). The focal sets are represented by different grey levels.

Fig. 16. Segmentation of the brain by combination of the partitions obtained from individual images. (a) pathology; (b) CSF and ventricles; (c) normal brain tissues. The lower approximations of the clusters are represented by light grey areas, the upper approximations by the union of light and dark grey areas.

Author biography

M.-H. Masson is engineer of the University of Technology of Compiègne and earned a doctorate in 1992 and a "Habilitation à diriger des Recherches"' in 2005 from the same institution. She is assistant Professor at the Université de Picardie Jules Verne, France and is a member of Heudiasyc Laboratory of the Université de Technologie de Compiègne, France since 1993. Her research interests include statistical pattern recognition, data analysis, uncertainty modelling and information fusion.

Thierry Denœux graduated in 1985 as an engineer from the Ecole Nationale des Ponts et Chaussées in Paris, and received a doctorate from the same institution in 1989. Currently, he is Full Professor with the Department of Information Processing Engineering at the Université de Technologie de Compiègne, France. His research interests concern belief functions theory, fuzzy data analysis and, more generally, the management of imprecision and uncertainty in data analysis, pattern recognition and information fusion. He is the Editor-in-Chief of the *International Journal of Approximate Reasoning*, and a member of the editorial board of *Fuzzy Sets and Systems*.