Belief C-Means: An Extension of Fuzzy C-Means Algorithm in Belief Functions Framework

Zhun-Ga Liu, Jean Dezert, Grégoire Mercier, Quan Pan

Belief C-Means: An extension of Fuzzy C-Means algorithm in belief functions framework

Zhun-ga Liu [a,c,*], Jean Dezert [b], Grégoire Mercier [c], Quan Pan [a]

[a] School of Automation, Northwestern Polytechnical University, Xi'an, China
[b] Onera – The French Aerospace Lab, F-91761 Palaiseau, France
[c] Télécom Bretagne, Technopôle Brest-Iroise 29238, France

The well-known Fuzzy C-Means (FCM) algorithm for data clustering has been extended to Evidential C-Means (ECM) algorithm in order to work in the belief functions framework with credal partitions of the data. Depending on data clustering problems, some barycenters of clusters given by ECM can become very close to each other in some cases, and this can cause serious troubles in the performance of ECM for the data clustering. To circumvent this problem, we introduce the notion of imprecise cluster in this paper. The principle of our approach is to consider that objects lying in the middle of specific classes (clusters) barycenters must be committed with equal belief to each specific cluster instead of belonging to an imprecise meta-cluster as done classically in ECM algorithm. Outliers object far away of the centers of two (or more) specific clusters that are hard to be distinguished, will be committed to the imprecise cluster (a disjunctive meta-cluster) composed by these specific clusters. The new Belief C-Means (BCM) algorithm proposed in this paper follows this very simple principle. In BCM, the mass of belief of specific cluster for each object is computed according to distance between object and the center of the cluster it may belong to. The distances between object and centers of the specific clusters and the distances among these centers will be both taken into account in the determination of the mass of belief of the meta-cluster. We do not use the barycenter of the meta-cluster in BCM algorithm contrariwise to what is done with ECM. In this paper we also present several examples to illustrate the interest of BCM, and to show its main differences with respect to clustering techniques based on FCM and ECM.

1. Introduction

In the data clustering analysis, the credal partition based on the belief functions theory has been introduced recently in (Denœux and Masson, 2003, 2004; Masson and Denœux, 2004, 2008). The credal partition is a general extension of the fuzzy (probabilistic) (Bezdek, 1981, 2000), possibilistic partition (Krishnapuram and Keller, 1996) and hard partition (Lloyd, 1982), and it allows the object not only to belong to single clusters, but also to belong to any subsets of the frame of discernment $\Omega = \{w_1, \ldots, w_c\}$ by allocating a mass of belief of each object to all elements of the power-set of Ω denoted 2^{Ω}. So the credal partitioning provides more refined partitioning results than the other partitioning techniques. This makes it very appealing for solving data clustering problems in practice.

The evidential clustering (EVCLUS) algorithm (Denœux and Masson, 2004) for relational data and the Evidential C-Means (ECM) (Masson and Denœux, 2008) for object data have been proposed originally by Denœux and Masson for the credal partitioning of data. In this paper, we focus on the problem of computing a credal partition from object data as in ECM context but using a different approach. ECM (Masson and Denœux, 2008) has been inspired from the Fuzzy C-Means (FCM) (Bezdek, 1981) and Dave's Noise-Clustering algorithm (Dave, 1991), and it can been seen as a direct extension of FCM in the belief functions framework. The mass of belief for each object is computed based on the distance between the object and the barycenters of focal elements that are subsets of Ω. The focal element composed by more than one singleton element of Ω is called an imprecise element and its corresponding cluster is called a meta-cluster. The cluster associated with a singleton element (a single class) is called a specific cluster (or a precise cluster). In ECM algorithm, the barycenter of a meta-cluster is obtained in averaging the centers of the specific clusters involved in the meta-cluster it is related with. It implies that the objects lying in the middle of the several specific clusters will be considered to belong to the meta-cluster represented by the union (disjunction) of these specific clusters. This way of processing is questionable because it can happen that the centers of different

* Corresponding author at: Télécom Bretagne, Technopôle Brest-Iroise 29238, France.
E-mail addresses: liuzhunga@gmail.com (Z.-g. Liu), jean.dezert@onera.fr (J. Dezert), Gregoire.Mercier@telecom-bretagne.eu (G. Mercier).

clusters are very close, and eventually that the centers overlap with each other, which is not efficient of course for data clustering with ECM. Thus, there is a serious difficulty for clustering the objects close to these similar/overlapped centers of meta-cluster and specific clusters.

For example, let's consider a set of data to be classified in three distinct classes $\Omega = \{w_1, w_2, w_3\}$ with the prototypes[1] $\mathbf{v_1}, \mathbf{v_2}$ and $\mathbf{v_3}$. In ECM, the center of the cluster $w_1 \cup w_3$ is given by $\mathbf{v_{1,3}} = \frac{\mathbf{v_1}+\mathbf{v_3}}{2}$, and the "ignorance center" is $\mathbf{v_\Omega} = \frac{\mathbf{v_1}+\mathbf{v_2}+\mathbf{v_3}}{3}$. However, if the centers of $\mathbf{v_2}$, $\mathbf{v_{1,3}}$ and $\mathbf{v_\Omega}$ are very close to each other, mathematically represented by $\mathbf{v_2} \approx \frac{\mathbf{v_1}+\mathbf{v_3}}{2}$, then $\mathbf{v_2} \approx \frac{\mathbf{v_1}+\mathbf{v_3}}{2} \approx \frac{\mathbf{v_1}+\mathbf{v_2}+\mathbf{v_3}}{3}$, the classification results about w_2, $w_1 \cup w_3$ and Ω will be difficult to be distinguished. Particularly, the data close to these centers can possibly be associated with the distinct cluster w_2, or with $w_1 \cup w_3$, or with Ω by ECM, and this seems not very reasonable.

In the new Belief C-Means (BCM) algorithm that we propose in this paper, the mass of belief of the specific cluster for each object is computed from the distance between the object and the center of the cluster, and the mass of belief of a meta-cluster is computed both from the distances between object and prototypes of the involved specific clusters, and the distances among these prototypes. In BCM, there is no need to compute the barycenter of the meta-clusters. At the end of this paper, we give some simple examples to show the interest of BCM with respect to FCM and ECM approaches.

2. Basics of Evidential C-Means (ECM)

ECM is a direct extension of FCM and it is based on a general model of partitioning called credal partitioning that refers to the framework of belief functions. The class membership of an object $\mathbf{x}_i = (x_{i_1}, \ldots, x_{i_p})$ is represented by a bba $m_i(.)$ over a given frame of discernment $\Omega = \{w_1, \ldots, w_c\}$, where $|\Omega| = c$ is known. $p \geqslant 1$ is the dimension of the attribute vector $\mathbf{x}_i$ associated with the ith object. This representation is able to model all situations ranging from complete ignorance to full certainty concerning the class of $\mathbf{x}_i$. In ECM, the mass of belief for associating the object $\mathbf{x}_i$ with an element A_j of 2^Ω denoted by $m_{ij} \triangleq m_{\mathbf{x}_i}(A_j)$, is determined from the distance d_{ij} between $\mathbf{x}_i$ and the prototype vector $\bar{\mathbf{v}}_j$ of the element A_j. Note that A_j can either be a single class, an union of single classes, or the whole frame Ω. The prototype vector $\bar{\mathbf{v}}_j$ of A_j, is defined as the mean vector of the prototype attribute vectors of the singletons of Ω included in A_j. $\bar{\mathbf{v}}_j$ is defined mathematically by

$$\bar{\mathbf{v}}_j = \frac{1}{c_j} \sum_{i=1}^{c} s_{kj} \mathbf{v}_k \quad \text{with} \quad s_{kj} = \begin{cases} 1, & \text{if } w_k \in A_j, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $\mathbf{v}_k$ is the prototype attribute vector of (i.e. the center of the single cluster associated with) the single class w_k, and $c_j = |A_j|$ denotes the cardinality of A_j, and d_{ij} is defined by:

$$d_{ij} = \|\mathbf{x}_i - \bar{\mathbf{v}}_j\|^2, \tag{2}$$

where $\|\mathbf{z}\| = \sqrt{z_1^2 + z_2^2 + \cdots + z_n^2}$ denotes the Euclidean norm of a n-dimensional vector.

In ECM, the determination of $m_{ij} \triangleq = m_{\mathbf{x}_i}(A_j)$ from d_{ij} is done in such a way that m_{ij} is low (resp. high) when d_{ij} is high (resp. low). Actually, m_{ij} is obtained by the minimization of the following objective function under a constraint to obtain the best credal partitioning problem (see Masson and Denœux, 2008 for justifications and details):

$$J_{ECM} = \sum_{i=1}^{n} \sum_{A_j \subseteq \Omega, A_j \neq \emptyset} c_j^\alpha m_{ij}^\beta d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^\beta. \tag{3}$$

[1] A prototype is a typical attribute vector characterizing a class. Usually the prototype is chosen as the center of the given class under consideration.

Because m_{ij} must be a basic belief assignment, the following constraint must be satisfied for any object $\mathbf{x}_i$

$$\sum_{A_j \subseteq \Omega, A_j \neq \emptyset} m_{ij} + m_{i\emptyset} = 1 \tag{4}$$

The solution of the minimization of (3) under the constraint (4) has been established by Masson and Denœux (2008) and it is given for each object $\mathbf{x}_i$, $(i = 1, 2, \ldots, n)$ by:

- For all $A_j \subseteq \Omega$ and $A_j \neq \emptyset$,

$$m_{ij} = \frac{c_j^{-\alpha/(\beta-1)} d_{ij}^{-2/(\beta-1)}}{\sum_{A_k \neq \emptyset} c_k^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)} + \delta^{-2/(\beta-1)}}, \tag{5}$$

where α is a tuning parameter allowing to control the degree of penalization; β is a weighting exponent (its suggested default value in (Masson and Denœux, 2008) is $\beta = 2$); δ is a given threshold tuning parameter for the filtering of the outliers; $c_j = |A_j|$ is a weighting coefficient for penalizing the subsets with high cardinality.
- For $A_j = \emptyset$,

$$m_{i\emptyset} \triangleq m_{\mathbf{x}_i}(\emptyset) = 1 - \sum_{A_j \neq \emptyset} m_{ij}. \tag{6}$$

The centers of the class are given by the rows of the matrix $V_{c \times p}$

$$V_{c \times p} = H_{c \times c}^{-1} . B_{c \times p}, \tag{7}$$

where the elements B_{lq} of $B_{c \times p}$ matrix for $l = 1, 2, \ldots, c$, $q = 1, 2, \ldots, p$, and the elements H_{lk} of $H_{c \times c}$ matrix for $l, k = 1, 2, \ldots, c$ are given by:

$$B_{lq} = \sum_{i=1}^{n} x_{i_q} \sum_{w_l \in A_j} c_j^{\alpha-1} m_{ij}^\beta, \tag{8}$$

$$H_{lk} = \sum_{i=1}^{n} \sum_{\{w_k, w_l\} \subseteq A_j} c_j^{\alpha-2} m_{ij}^\beta. \tag{9}$$

3. Belief C-Means (BCM) approach

3.1. Basic principle of BCM

In ECM, the prototype vector (i.e. the center) of an imprecise (i.e. a meta) cluster is obtained by averaging the prototype vectors of the specific clusters included in it, as shown in (1). ECM method is of course relatively easy to apply, but it yields to serious problems in some difficult cases of data clustering where the prototype vectors of the specific clusters overlap with the meta-clusters. This problem will cause troubles in the association of an object with a particular specific cluster or the meta-cluster the object may also belong to. That is why a better approach must be developed to circumvent this problem. This is the purpose of our BCM algorithm.

In BCM approach, we consider that when a data belongs to a meta-cluster (i.e. to an imprecise class corresponding to the disjunction of several single classes), this means that the prototypes of the single classes in the meta-cluster are quite difficult to be distinguished (discerned) from the object under analysis. More clearly, if the prototype vectors of the classes included in a given meta-cluster are close to each other and in the meantime they are far from the object attribute vector, then it seems more reasonable and natural to commit this object rather to the meta-cluster, than to each of these specific classes as if they were considered separately.

To illustrate this very reasonable BCM principle, let's consider only two objects $\mathbf{x}_1$ and $\mathbf{x}_2$ and three possible centers of clusters (prototypes) $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ corresponding to the classes w_1, w_2 and w_3 as shown in Fig. 1.

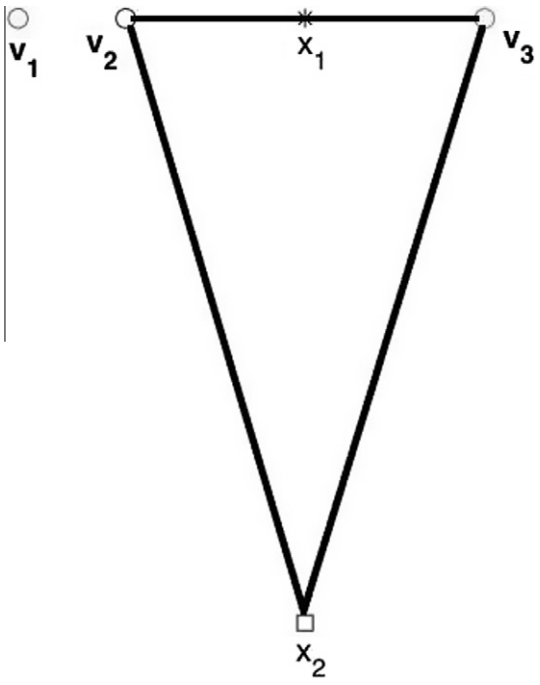


Fig. 1. Uncertainty versus imprecision for object association.

Let's explain the main differences between the different approaches from this simple figure:

- In the FCM approach, the objects $\mathbf{x}_1$ and $\mathbf{x}_2$ are considered with equal membership for classes w_2 and w_3 because $\mathbf{x}_1$ is at the same distance with respect to $\mathbf{v}_2$ and to $\mathbf{v}_3$, and because $\mathbf{x}_2$ is also at the same distance with respect to $\mathbf{v}_2$ and to $\mathbf{v}_3$, even if the distance of $\mathbf{x}_2$ to $\mathbf{v}_2$ (resp. to $\mathbf{v}_3$) is much bigger than the distance of $\mathbf{x}_1$ to $\mathbf{v}_2$ (resp. to $\mathbf{v}_3$).
- In ECM approach, $\mathbf{x}_1$ will likely be committed to the meta-cluster $w_2 \cup w_3$ because it is in the middle of the centers (prototypes) of w_2 and w_3, whereas $\mathbf{x}_2$ will likely be committed to the separate classes w_2 and w_3 with same mass of belief or the meta-cluster $w_2 \cup w_3$ depending on the tuning of the parameter α in ECM.
- In BCM approach and contrariwise to ECM, we consider that $\mathbf{v}_2$ and $\mathbf{v}_3$ are clearly distinct from the object $\mathbf{x}_1$ position, and therefore the object $\mathbf{x}_1$ should better be committed with equal masses of belief to the separate classes w_2 and w_3 rather than to the meta-cluster $w_2 \cup w_3$. Conversely, the prototypes $\mathbf{v}_2$ and $\mathbf{v}_3$ are more difficult[2] to be discerned from the object $\mathbf{x}_2$, since the distance between $\mathbf{x}_2$ and $\mathbf{v}_2$ or $\mathbf{v}_3$ is much larger than the distance between $\mathbf{v}_2$ and $\mathbf{v}_3$. Moreover, the object $\mathbf{x}_2$ is farther to the prototype $\mathbf{v}_1$ than to prototypes $\mathbf{v}_2$ and $\mathbf{v}_3$. So, it seems reasonable for the object $\mathbf{x}_2$ to commit more mass of belief to the meta-cluster $w_2 \cup w_3$ than separately to classes w_2 and w_3 in such conditions. So we see that the principle of BCM is in opposition with ECM principle for determining the basic belief assignments and to find the credal partition of objects with specific clusters, or meta-clusters.

In summary, in BCM the mass of belief committed to a singleton class for an object will depend on the distance between the object and the center of the specific cluster of the class, whereas the mass of belief committed to a meta-cluster will depend on the distances between the object and the prototypes of the specific clusters belonging to the meta-cluster, as well as on the distances between these prototypes.

3.2. The BCM algorithm

Let's present here the BCM approach in a general context. We consider a set of $n \geqslant 1$ objects. Each object #i (called also a data point) is represented by a given attribute vector $\mathbf{x}_i$ of dimension $p \geqslant 1$. These objects must be classified into $k \geqslant 2$ classes over a given frame $\Omega = \{w_1, w_2, \ldots, w_k\}$ with the corresponding centers $\{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_k}\}$. Each center $\mathbf{v_k}$ corresponds actually to the prototype of the class w_k. Like in FCM and ECM approaches, in BCM approach the mass of belief (bba) $m_{\mathbf{x}_i}(w_j)$ of $\mathbf{x}_i$ committed to the class w_j, is assumed to increase with the decrease of the distance $d_{\mathbf{x_i v_j}}$ between $\mathbf{x}_i$ and the center $\mathbf{v_j}$. The smaller $d_{\mathbf{x_i v_j}}$ leads to the bigger $m_{\mathbf{x}_i}(w_j)$. If the object $\mathbf{x}_i$ is closer[3] to the centers $\mathbf{v_j}, \mathbf{v_{j+1}}, \ldots, \mathbf{v_t}$ for some j, $t < k$, and in the meanwhile the distances $d_{\mathbf{v_j v_{j+1}}}, \ldots, d_{\mathbf{v_j v_t}}, \ldots, d_{\mathbf{v_{t-1} v_t}}$ between the centers of the clusters is smaller,[4] then the value committed to $m_{\mathbf{x}_i}(w_j \cup w_{j+1} \ldots \cup w_t)$ by BCM will naturally increase. In BCM approach, we propose to compute $m_{\mathbf{x}_i}(w_j \cup w_{j+1} \cup \ldots w_t)$ as the simple average of the distances $d_{\mathbf{x_i v_j}}, \ldots, d_{\mathbf{x_i v_t}}$ and $d_{\mathbf{v_j v_{j+1}}}, \ldots, d_{\mathbf{v_j v_t}}, \ldots, d_{\mathbf{v_{t-1} v_t}}$. If some few objects (data points) are far from all the centers of clusters, they will be considered as outliers and committed to $\emptyset$ (or to an extra class if one prefers).

It is clear that we use here in our BCM approach a simple average principle to keep algorithm as simple as possible, but more advanced techniques methods to compute bba committed to partial ignorance from $\{d_{\mathbf{x_i v_j}}\}$ and $\{d_{\mathbf{v_i v_p}}\}$ could be used instead, but the optimization problem will appear much harder to solve. Advanced methods to compute bba is out of the scope of this paper and they will be investigated in future research works. In our BCM approach, the objective function J_{BCM} that we propose to minimize, under the same constraint (4) as for ECM, differs of J_{ECM} objective function (3). It is defined as

$$J_{BCM}(M,V) = \Delta_1 + \Delta_2 + \Delta_3 \tag{10}$$

with

$$\begin{cases} \Delta_1 = \sum_{i=1}^{n} \sum_{A_j \in \Omega, |A_j|=1} c_j^{\alpha} m_{ij}^{\beta} d_{ij}^2, \\ \Delta_2 = \sum_{i=1}^{n} \sum_{A_j \in \Omega, |A_j|>1} c_j^{\alpha} m_{ij}^{\beta} \overline{d_{ij}^2}, \\ \Delta_3 = \sum_{i=1}^{n} \delta^2 m_{i\phi}^{\beta}. \end{cases} \tag{11}$$

and

$$\overline{d_{ij}^2} = \frac{\sum_{w_k \in A_j} d_{ik}^2 + \sum_{w_x, w_y \in A_j} \gamma d_{xy}^2}{|A_j| + \gamma C_{|A_j|}^2}. \tag{12}$$

The tuning parameters α, β, δ and $c_j = |A_j|$ have the same meaning as in ECM; γ is the weighting factor of the distances among the centers; $d_{ij} \triangleq d_{\mathbf{x_i v_j}}$ is the distance between the data point $\mathbf{x}_i$ and the class w_j; and $d_{xy} \triangleq d_{\mathbf{v_x v_y}}$ is the distance between the classes w_x and w_y. $C_{|A_j|}^2 = \frac{|A_j|!}{2!(|A_j|-2)!}$ is the number of combinations of $|A_j|$ taken 2 at a time.

The quantities Δ_1, Δ_2 and Δ_3 entering in J_{BCM} objective function are well justified since:

[2] To easily understand this, we can imagine that the object $\mathbf{x}_2$ is very far away from $\mathbf{v}_2$ and $\mathbf{v}_3$ and located on the line perpendicular to the line $\mathbf{v}_2$ and $\mathbf{v}_3$ line and passing through $(\mathbf{v}_2 + \mathbf{v}_3)/2$.

[3] This indicates that $\mathbf{x}_i$ has potentially more chance to belong to the classes $v_j, v_{j+1}, \ldots, v_t$ than to other classes.

[4] This indicates that the precise class is more difficult to be discerned from x_i.

- Δ_1 indicates that the belief of an object (a data point) associated to the singleton cluster is proportional to the distance between the data point and the center of the single cluster.
- Δ_2 indicates that the belief of a data point associated to an imprecise cluster is proportional to the average distance between the data point and the centers involved in the imprecise cluster and also to the distances between the centers of the elementary clusters. The weight of the distances between the centers of the elementary clusters can be tuned according to the applications in BCM.
- Δ_3 indicates that if a data point is very far (according to a predetermined threshold δ) from all the centers of elementary clusters, then this data will be considered as outlier represented by the $\emptyset$ "class".

In the J_{BCM} objective function (10), Δ_1 and Δ_2 can be actually summed into one single term as follows:

$$\Delta_{12} = \Delta_1 + \Delta_2 = \sum_{i=1}^{n} \sum_{A_j \in \Omega} c_j^{\alpha} m_{ij}^{\beta} \overline{d_{ij}^2}, \tag{13}$$

$$\text{where } \overline{d_{ij}^2} = \frac{\sum_{w_k \in A_j} d_{ik}^2 + \sum_{w_x, w_y \in A_j} \gamma d_{xy}^2}{|A_j| + \gamma C_{|A_j|}^2}. \tag{14}$$

As with ECM, several tuning parameters have to be set before using BCM algorithm. For example, the number of classes c can be determined by minimizing the validity index of a credal partition as the average normalized specificity proposed in (Masson and Denœux, 2008):

$$N^*(c) = \frac{1}{n \log_2(c)} \sum_{i=1}^{n} \left[\sum_{A \in 2^{\Omega} \setminus \emptyset} m_i(A) \log_2 |A| + m_i(\emptyset) \log_2(c) \right], \tag{15}$$

where $1 \leqslant N^*(c) \leqslant 1$.

As in FCM (Bezdek, 1981), PCM (Krishnapuram and Keller, 1996) or ECM (Masson and Denœux, 2008) approaches, $\beta = 2$ can be used as default value, and we used this value in our simulations presented in the sequel. α allows to control the number of points assigned to focal elements of high cardinality, and it can be tuned according to actual applications: the higher α, the less imprecise will be the resulting partition. δ is the threshold of the outliers, and it is strongly data-dependent. If most data is very near to the corresponding center, δ can be small. γ plays an important role in BCM. If γ is too large as $\gamma > 1$, the distances among the prototypes will take the bigger weight than the distance between the data and the prototypes. It implies that even the average distances between the data and the prototypes is smaller than the average distances among these prototypes, the data are still likely to be classified into meta-cluster. So the higher of γ will lead a more imprecise credal partition. If γ is too small, some data far from the others will still be classified into specific clusters as the other objects, which betrays the basic idea we follow in BCM. γ should be tuned according to the actual application, and it is generally suggested to be taken in $(0, 0.7)$.

To implement BCM algorithm, we need to minimize J_{BCM} criterion and propose a decision-making support.

3.2.1. Minimization of J_{BCM}

To minimize J_{BCM}, we use Lagrange multipliers method. In the first step, the centers of the clusters V are considered fixed. Lagrange multipliers λ_i are used to solve the constrained minimization problem with respect to M as follows:

$$\mathcal{L}(M, \lambda_1, \ldots, \lambda_n) = J_{BCM}(M, V) - \sum_{i=1}^{n} \lambda_i \left(\sum_{j / A_j \subseteq \Omega A_j \neq \emptyset} m_{ij} + m_{i\emptyset} - 1 \right). \tag{16}$$

M is a matrix composed by the bba m_{ij}, where the number of the objects $i = 1, 2, \ldots, n$, and the number of the focal elements $j = 1, 2, \ldots, 2^{|\Omega|}$.

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

$$\frac{\partial \mathcal{L}}{\partial m_{ij}} = c_j^{\alpha} \beta m_{ij}^{\beta - 1} \overline{d_{ij}^2} - \lambda_i = 0, \tag{17}$$

$$\frac{\partial \mathcal{L}}{\partial m_{i\emptyset}} = \beta m_{i\emptyset}^{\beta - 1} \delta^2 - \lambda_i = 0, \tag{18}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = \sum_{j / A_j \subseteq \Omega A_j \neq \emptyset} m_{ij} + m_{i\emptyset} - 1 = 0. \tag{19}$$

We thus have from (17)

$$m_{ij} = \left(\frac{\lambda_i}{\beta} \right)^{\frac{1}{\beta - 1}} \left(\frac{1}{c_j^{\alpha} \overline{d_{ij}^2}} \right)^{\frac{1}{\beta - 1}} \tag{20}$$

and from (18)

$$m_{i\emptyset} = \left(\frac{\lambda_i}{\beta} \right)^{\frac{1}{\beta - 1}} \left(\frac{1}{\delta^2} \right)^{\frac{1}{\beta - 1}}, \tag{21}$$

using (17)–(19)

$$\left(\frac{\lambda_i}{\beta} \right)^{1/(\beta - 1)} = \frac{1}{\sum_j \frac{1}{c_j^{\alpha/(\beta-1)} \overline{d_{ij}^{2/(\beta-1)}}} + \frac{1}{\delta^{2/(\beta-1)}}}. \tag{22}$$

Returning in (17), one obtains the necessary condition of optimality for M:

$$m_{ij} = \frac{c_j^{-\alpha/(\beta-1)} \overline{d_{ij}^{-2/(\beta-1)}}}{\sum_{A_k \neq \emptyset} c_k^{-\alpha/(\beta-1)} \overline{d_{ik}^{-2/(\beta-1)}} + \delta^{-2/(\beta-1)}}, \tag{23}$$

$$m_{i\emptyset} = 1 - \sum_{A_j \neq \emptyset} m_{ij}, \quad \forall i = 1, \ldots, n. \tag{24}$$

Note that these update equations are very similar to those of the ECM algorithm except that the distances among the centers are considered here and there is no need to compute the centers of meta-clusters.

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

$$\frac{\partial J_{BCM}}{\partial \mathbf{v_l}} = \sum_{i=1}^{n} \sum_{A_l \subseteq A_j} c_j^{\alpha} m_{ij}^{\beta} \frac{\partial \overline{d_{ij}^2}}{\partial \mathbf{v_l}} \tag{25}$$

with

$$\frac{\partial \overline{d_{ij}^2}}{\partial \mathbf{v_l}} = \frac{-2(\mathbf{x_i} - \mathbf{v_l}) + 2\gamma \sum_{A_y \in A_j} (\mathbf{v_l} - \mathbf{v_y})}{|A_j| + \gamma C_{|A_j|}^2}. \tag{26}$$

Thus,

$$\frac{\partial J_{BCM}}{\partial \mathbf{v_l}} = \sum_{i=1}^{n} \sum_{A_l \subseteq A_j} c_j^{\alpha} m_{ij}^{\beta} \frac{-2(\mathbf{x_i} - \mathbf{v_l}) + 2\gamma \sum_{A_y \in A_j} (\mathbf{v_l} - \mathbf{v_y})}{|A_j| + \gamma C_{|A_j|}^2}. \tag{27}$$

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

Table 1
Belief C-Means algorithm.

Input:	Data to cluster: $X_1,\ldots,X_n$ in $\mathbb{R}^p$
Parameters:	c: number clusters, $2 \leqslant c < n$ $\alpha \geqslant 0$: weighting exponent for cardinality $\beta > 1$: weighting exponent $\delta > 0$: distance for outlier ($\emptyset$-"class") $\epsilon > 0$: termination threshold $\gamma > 0$: weight of distances among centers
Initialization:	Choose randomly initial mass M_0 $t \leftarrow 0$ Repeat $t \leftarrow t + 1$ Compute B_t and H_t by (30)–(32),; Compute V_t by solving (29); Compute M_t using (23), (24); until $\|V_t - V_{t-1}\| < \epsilon$

$$\sum_{i=1}^{n}\sum_{A_l\subseteq A_j}\frac{2c_j^{\alpha}m_{ij}^{\beta}\mathbf{x}_i}{|A_j|+\gamma C_{|A_j|}^2}=\sum_{i=1}^{n}\sum_{A_l\subseteq A_j}\times\frac{2c_j^{\alpha}m_{ij}^{\beta}\left[(\gamma|A_j|-\gamma+1)\mathbf{v_l}-2\gamma\sum\limits_{A_y\in A_j}\mathbf{v_y}\right]}{|A_j|+\gamma C_{|A_j|}^2}. \tag{28}$$

The system of linear equations can be equally represented by:

$$B_{c\times n}X_{n\times p}=H_{c\times c}V_{c\times p}, \tag{29}$$

where

$$B_{lh}=\frac{\sum_{A_l\in A_j}c_j^{\alpha}m_{hj}^{\beta}}{|A_j|+\gamma C_{|A_j|}^2}, \tag{30}$$

$$H_{ll}=\sum_{i=1}^{n}\sum_{A_l\in A_j}\frac{c_j^{\alpha}m_{ij}^{\beta}(\gamma|A_j|-\gamma+1)}{|A_j|+\gamma C_{|A_j|}^2}, \tag{31}$$

$$H_{lq}=-\sum_{i=1}^{n}\sum_{A_l,A_q\in A_j}\frac{c_j^{\alpha}m_{ij}^{\beta}\gamma}{|A_j|+\gamma C_{|A_j|}^2},\quad l\neq q. \tag{32}$$

V is the solution of the above linear system (29), and it can be solved by using a standard linear system solver. It is worth to note that $H_{c\times c}$ possibly is a matrix without full rank in few particular cases. If so, we can make $H_{c\times c}$ be a full rank matrix with the rank equal to c by the minor tuning of the parameter γ. The pseudo-code of the BCM algorithm is given in Table 1 for convenience.

Remark. The initial bba M_0 can be randomly generated over 2^{Θ} (the empty set can also have a positive mass in the initial choice of the bba) but the final clustering results are not very sensitive to the choice of the initialization of M_0 after the process of optimization.

3.2.2. Decision-making support

For decision-making support with hard partition, belief function *Bel*(.), or plausibility function *Pl*(.) (Shafer, 1976), or pignistic probability *BetP*(.) (Smets, 2005; Smets and Kennes, 1994; Smets, 1990) are common chosen.[5] The belief, plausibility functions and the pignistic transformation are introduced as follows:

$$Bel(A)=\sum_{A,B\in 2^{\Omega};B\subseteq A}m(B), \tag{33}$$

$$Pl(A)=\sum_{A,B\in 2^{\Omega};A\cap B\neq\emptyset}m(B). \tag{34}$$

The interval $[Bel(A),Pl(A)]$ is then interpreted as the lower and upper bounds of imprecise probability for decision-making support (Shafer, 1976) and the pignistic probability $BetP(A)$ commonly used to approximate the unknown probability $P(A)$ in $[Bel(A),Pl(A)]$ is calculated by:

$$BetP(A)=\sum_{A,B\in 2^{\Omega},A\subseteq B}\frac{|A\cap B|}{|B|}m(B), \tag{35}$$

where $|X|$ is the cardinality of the element X.

4. Examples

Example 1. The Diamond data set:We illustrate the behavior of BCM and show the difference of BCM with respect to ECM from an example similar to the classical data set (Windham, 1985). Objects 1 and 11 are part of Windham's data whereas object 13 is an outlier. Object 12 is not close to the objects 1 and 11, but it is not so far to them as the outlier as shown in Fig. 2(a). The result of FCM obtained by the maximal fuzzy membership is shown in Fig. 2(b). Because the result produced by FCM is in the probability framework, there is no imprecise meta-cluster involved in FCM. Therefore, all the data including the objects 12 and 13 are just classified into w_1 or w_2 with the fuzzy membership about w_1 and w_2 as shown in Fig. 2(c).

In our simulation, BCM and ECM run with the same tuning parameters: $\alpha = 1/6$, $\beta = 2$, $\delta^2 = 125$, $\epsilon = 10^{-3}$, and in BCM one has taken $\gamma = 0.6$. A 2-credal partition was imposed so that the four focal elements have been considered in the optimization process as: w_1, w_2, Ω, and $\emptyset$ for representing the unknown extra class. The masses $m(w_1)$, $m(w_2)$, $m(\Omega)$ and $m(\emptyset)$ are shown in Fig. 3(b) and Fig. 4(b). The figures Fig. 3(a) and Fig. 4(a) show the results of clustering obtained by ECM and BCM respectively based on the maximal mass of belief.

In ECM and BCM, the classified results are extended in the belief functions framework with the meta-clusters. The points indexed by 1–5 and 8–11 are classified into two specific clusters by both ECM and BCM, and the point 13 is considered as an outlier. Most mass of belief is focused on ignorance Ω for point 6 and point 7 by ECM as shown in Fig. 3(b), since they are near the center of Ω which is the average of the centers of w_1 and w_2. Whereas, the largest mass of belief is respectively committed to w_1 and w_2 for points 6 and 7 by BCM as in Fig. 4(b). This is because the prototypes of w_1 and w_2 are distinct for points 6 and 7, and the point 6 is closer to the prototype of w_1 while point 7 is closer to the prototype of w_2. Therefore, the point 6 belongs to w_1, and the point 7 belongs to w_2. The point 12 is committed to the same meta-cluster by both ECM and BCM, but they follow two totally different principles. In ECM, it is because the point 12 is close to the center of Ω. Whereas in BCM, it is considered that the two prototypes w_1 and w_2 are hard to be distinguished for the point 12, and the point 12 is not close to the prototypes of w_1 and w_2. So it is also committed to Ω by BCM. Moreover, we can observe that Ω acquires a bit more mass of belief for the point 12 with BCM than with ECM when comparing Fig. 3(b) with Fig. 4(b). The point 13 is too far from the prototypes, and it is of course classified as outlier (i.e. $\emptyset$-"class") both with ECM and BCM. The difference of the classification results between ECM and BCM mainly lies in the different interpretations of the meta-clusters.

[5] *DSmP*(.) transformation proposed in Smarandache and Dezert (2004-2009) that provides a better probabilistic informational content than *BetP*(.) can also be chosen instead. But *DSmP*(.) is more complicated to implement than *BetP*(.) and it has not been tested in our application for now.

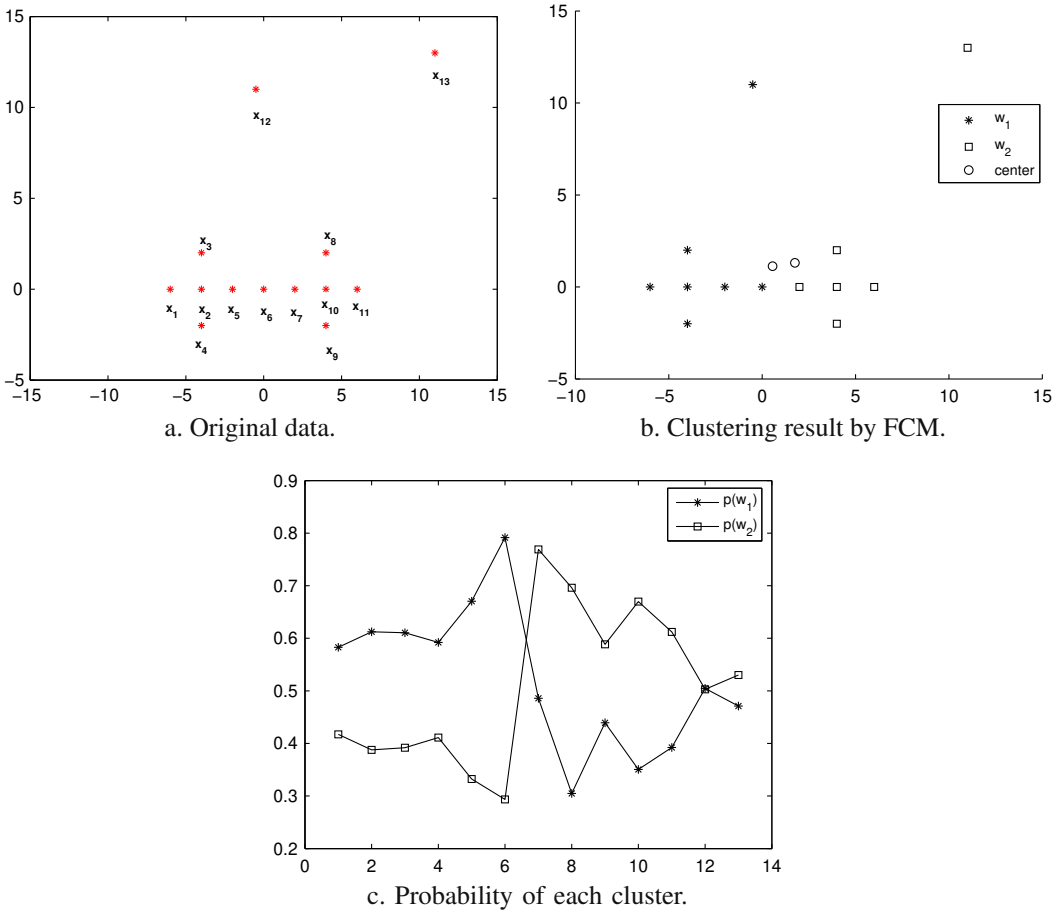


Fig. 2. Clustering of the data set by FCM.

Example 2. This example is designed to show the limitation of ECM. The artificial data are composed by $3 \times 100 + 10 = 310$ points. There are $2 \times 100 = 200$points generated around the centers $(0,0)$ and $(80,80)$ as Gaussian noise $\mathcal{N}(0,20)$, and 100 points are around $(160,160)$ with the Gaussian noise $\mathcal{N}(0,30)$, and 10 points lie around the center $(-130,310)$ with the Gaussian noise $\mathcal{N}(0,50)$. BCM and ECM have ben tested using with the same tuning parameters: $\alpha = 1/6$, $\beta = 2$, $\delta^2 = 90000$, and $\epsilon = 10^{-3}$. In BCM, one has taken $\gamma = 0.45$. The frequency of our CPU used for running our simulation code (developed under MatLab 7.0) was 2.0 GHz.

In this example, the data set is generated from the four centers and one subset of data is very small since it contains only 10 points, but the data of this subset have a very high of noise. So this small data set is truly a particular (very noisy) data set rather than the regular data set. In order to test in noisy conditions the BCM with the other methods, the number of the clusters for all the methods have been chosen $c = 3$.

The original data are shown in Fig. 5(a), and the classification results obtained with FCM by maximal fuzzy membership are shown in Fig. 5(b), and the results of the hard partition obtained by Lloyd (1982) are shown in Fig. 5(c). Fig. 5(d) and (e) show the results based on the maximal mass of belief obtained with ECM, and with BCM respectively.

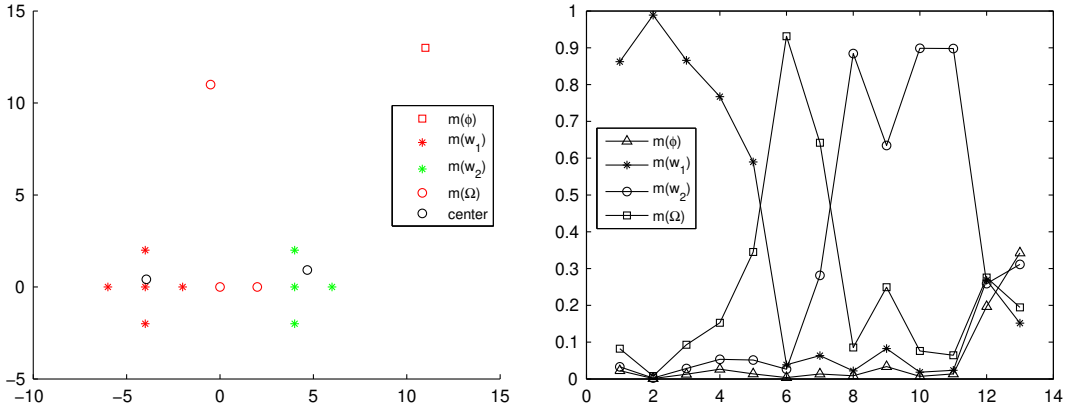
One can see on Fig. 5(d) that the set of the original data $\mathbf{x_2}$ (the set of original green points of Fig. 5(a)) is classified into even four meta-clusters as $\{w_1,w_2\}$, $\{w_1,w_3\}$, $\{w_2,w_3\}$ and the total ignorance Ω, and only a small part is covered by w_2 in ECM. It is mainly because the underlying centers of the four imprecise clusters lie around $(90,90)$, and they are close to each other, especially for the centers of $\{w_1,w_3\}$ and Ω. It implies that ECM is not suitable in such case because the centers of the meta-clusters overlap with those of the specific clusters. In FCM, the center of each cluster does not correspond well with the truth because of the influence of these 10 imprecise points, and the 10 points are all considered to be with class w_2. A part of points of $\mathbf{x_2}$ near the center of w_3 are classified to w_3, and another small part is classified to w_1. In BCM, most of the original data $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$ can be correctly classified. For the set of data $\mathbf{x_4}$ (i.e. the set of original black points of Fig. 5(a))), two points are considered as $\{w_1,w_2\}$, which indicates they cannot be distinguished by w_1 and w_2, and three points are classified into $\{w_2,w_3\}$. Two points are committed to ignorance, which means they are indistinguishable for the three classes. The other three points are too far from the prototypes, and they are considered as outliers. However, most of them are committed to w_2 by ECM.

Davies Bouldin index (DBI) (Davies and Bouldin, 1979) is commonly used as the clustering quality measure based on the hard partition decision result. This index represents the system-wide average of the similarity measures of each cluster with its most similar cluster, and the smaller DBI generally indicates the better clustering results. For the hard decision-making support, the mass of the belief of the clustering in ECM and BCM can be transformed into pignistic probability using (35) when a hard partition decision is necessary in some applications.

For example, if the bba associated with one point to be clustered is given by:

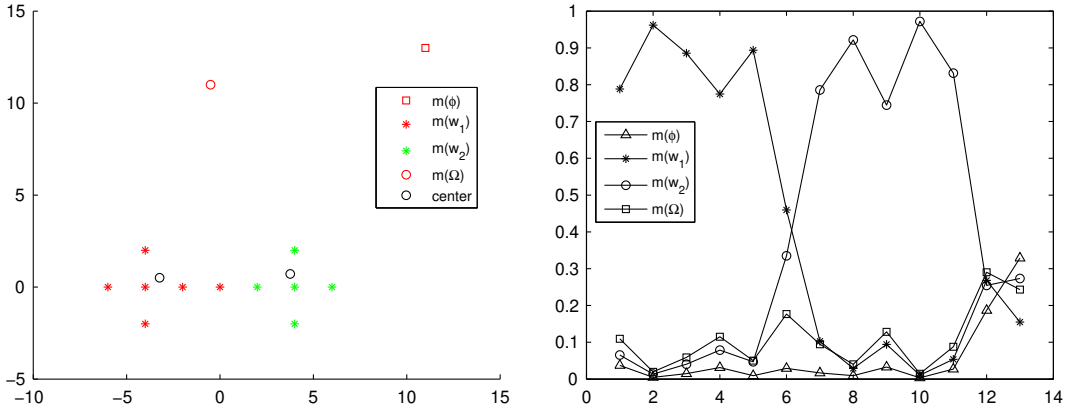
$$m(w_1) = 0.5, \quad m(w_2) = 0.2, \quad \text{and} \quad m(w_2,w_3) = 0.3.$$

then it can be transformed into a pignistic probability measure (Smets, 2005) as:



a. Clustering result by ECM. b. Mass of belief of each cluster.

Fig. 3. Clustering of the data set by ECM.



a. Clustering result by BCM. b. Mass of belief of each cluster.

Fig. 4. Clustering of the data set by BCM.

$$BetP(w_1) = m(w_1) = 0.5,$$

$$BetP(w_2) = m(w_2) + \frac{m(w_2, w_3)}{2} = 0.35,$$

$$BetP(w_3) = \frac{m(w_2, w_3)}{2} = 0.15.$$

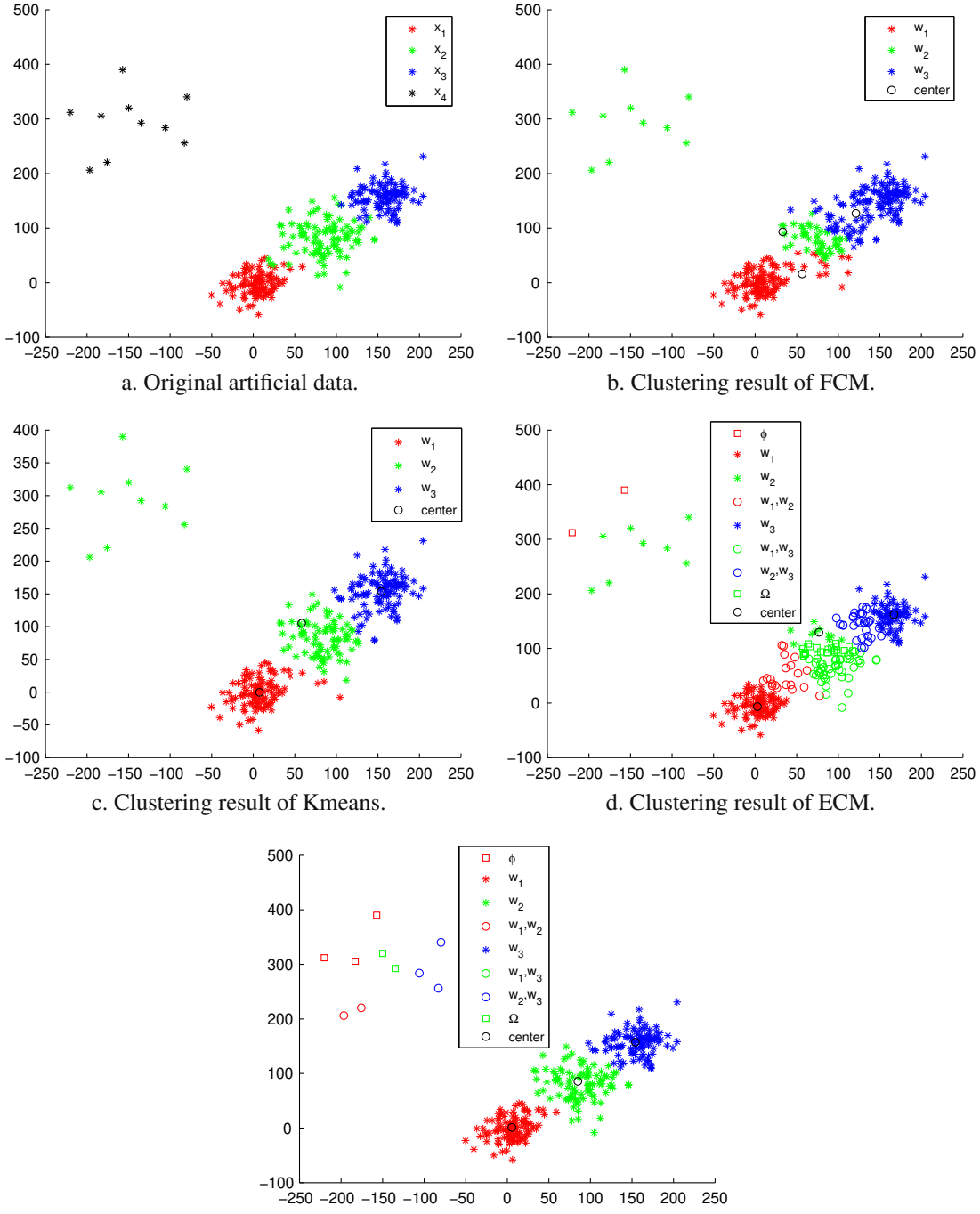
The hard decision of the cluster associated with this point is based on the maximal pignistic probability criterion. So in our eaxmple, the point will be clustered into w_1 since $max(BetP(.)) = BetP(w_1)$. DBI of the clustering results can be computed using this hard cluster decision. In our simulations, we have used $p = q = 2$ for computing DBI (see Davies and Bouldin, 1979 for details). DBI and execution time of the different methods are given in Table 2.

As we can see, BCM and ECM take a longer execution time than FCM and K means, and this is the price one has to pay for acquiring the credal partition. Moreover, ECM takes more time than BCM in this test, since it has been observed that ECM requires a greater number of iterations than BCM for completing the optimization process (using the same termination threshold). DBI of BCM is similar to that of FCM and K means, and is a bit bigger than that of ECM. It indicates the hard partition of ECM is the best and BCM produces better hard partition results than that of FCM and K means from the point view of DBI. Nevertheless, ECM provides very unreasonable credal partition. BCM produces more reasonable classification results using the belief functions framework, which is potentially very interesting in some applications.

Example 3. The iris flower data set (Fisher, 1936) is a typical test case for data clustering and classification, and it is used in this third experiment. The data set consists of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor), and the number of the total samples is $3 \times 50 = 150$. Four features were measured from each sample, they are the length and the width of sepal and petal, in centimeters. In order to show the effectiveness of BCM for dealing with the noisy data, five additional artificial points are added in the test data set as follows: X=[6 13 10 2; 7 6 4 3; 3 12 11 0.2; 6 8 8 0.2; 1 0.3 8 10]. The parameters used in BCM and ECM are: $\alpha = 1/6$, $\beta = 2$, $\delta^2 = 100$, $\epsilon = 10^{-3}$, and in BCM one has taken $\gamma = 0.42$.

The original data are shown by Fig. 6(a), and the classification results of FCM, K means, ECM and BCM with same criterion as in Example 2 are respectively shown by Fig. 6(b)–(e).

One can see that the classification results of FCM, K means and BCM are similar for the original Iris data except the additional artificial data, and most of the data are correctly classified. Nevertheless, a lot of the original data close to the middle of the clustering prototypes are considered as imprecise meta-cluster by ECM, which is unreasonable, and it is difficult to get the specific and correct classification results. Now let us focus on the five additional artificial data which is far from the other data, they are all clustered into w_1 by FCM and K means, since FCM and K means work in the probability framework, and they cannot provide more useful (specific) clustering information as the credal partition. Three of



a. Original artificial data.

b. Clustering result of FCM.

c. Clustering result of Kmeans.

d. Clustering result of ECM.

e. Clustering result of BCM.

Fig. 5. Clustering of artificial the data set by different methods.

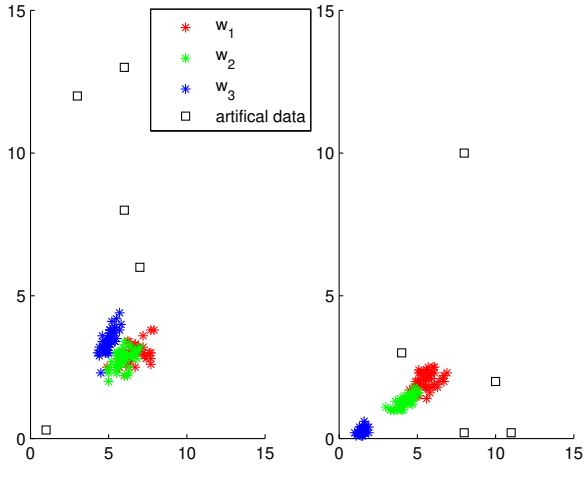
artificial data quite far from the other data are considered as outliers by both ECM and BCM. The other two data which are not far enough from the others are classified as $\{w_1, w_2\}$ by BCM, since they are more close to w_1 and w_2 than to w_3. This clustering results especially for the additional artificial data illustrate the interest and effectiveness of this new BCM approach.

DBI of the clustering results can still be obtained by the maximal pignistic probability criterion. DBI with the parameters $p = q = 2$ and execution time of the different methods are given in Table 3.

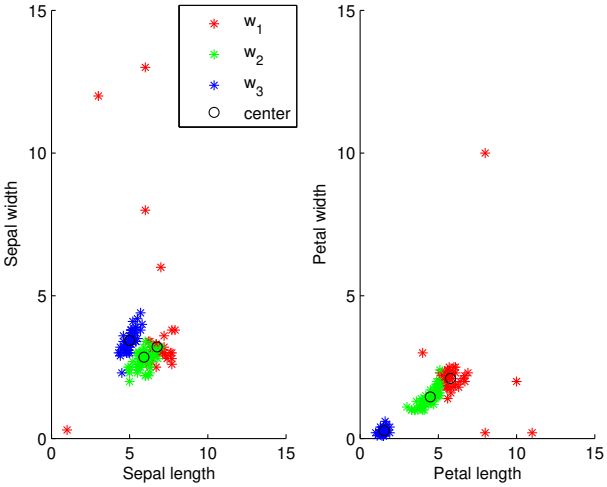
One can observe that the cost of execution time for BCM and ECM is much larger than for FCM and K means, since BCM and ECM produce the credal partitions which is more general than a fuzzy partition. We can see moreover that ECM still takes longer time than BCM because of the number of iteration in optimization process. DBI of BCM seems similar to that of FCM and it is a bit bigger than that of ECM, but the credal partitions of ECM is obviously worse than that of BCM. It shows the effectiveness of the proposed method BCM.

Table 2
DBI and execution time of different methods tested on artificial data.

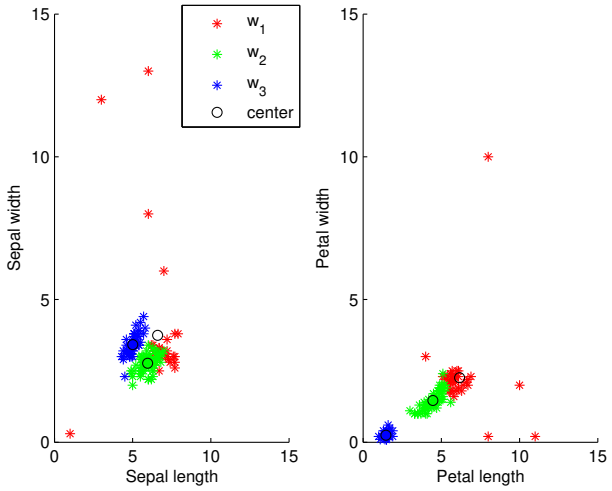
	K means	FCM	BCM	ECM
Execution time (in sec)	0.0156	0.0468	0.6680	0.9961
DBI	0.8080	0.8049	0.7712	0.6581



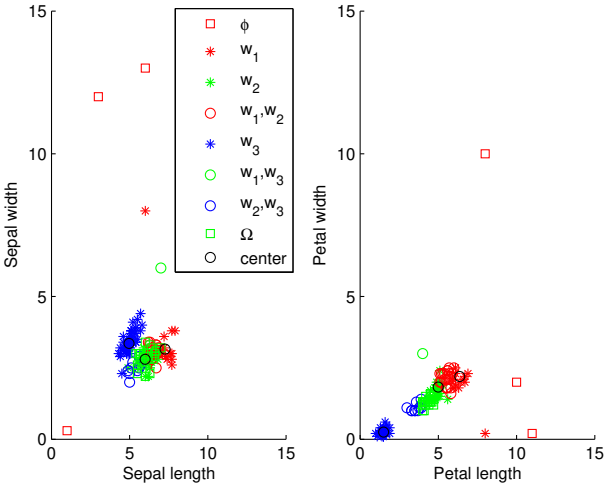
a. Iris data.



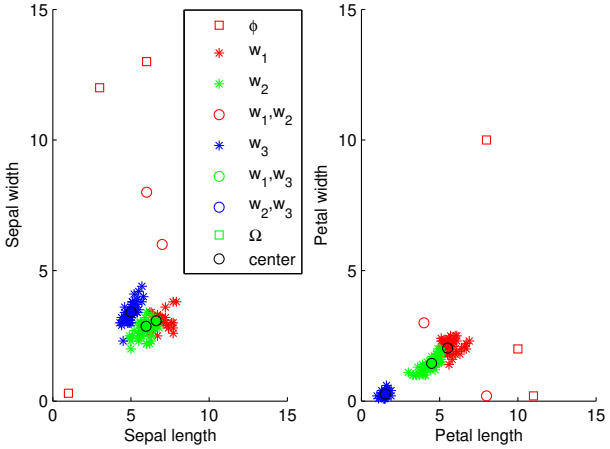
b. Clustering result of FCM.



c. Clustering result of Kmeans.



d. Clustering result of ECM.



e. Clustering result of BCM.

Fig. 6. Clustering of the Iris data set by different methods.

5. Conclusion

A new unsupervised clustering method, called BCM (Belief C-means), has been proposed and evaluated in this paper. BCM is an extension of FCM and an alternative of ECM. After analyzing the limitation of ECM, we have proposed another interpretation of meta-cluster defining bba's. In the determination of a mass of belief associated with the meta-cluster, not only distances between object and prototypes of specific clusters involved in the meta-cluster does count, but also the distances between those proto-

Table 3
DBI and execution time of different methods for Iris data.

	K means	FCM	BCM	ECM
execution time (in sec)	0.0312	0.0624	0.4588	0.5304
DBI	0.9723	0.8450	0.8811	0.6778

types. If an object is far from the prototype of a specific cluster compared with the distances among these prototypes, but it is still below the threshold of outlier filtering this object, it will be considered more possible to be committed to a meta-cluster composed by those specific classes. The mass of belief associating an object with specific cluster is computed according to the distance between the object and the prototypes of clusters like in ECM. Some examples have been given to illustrate the interest of BCM and to show its difference with respect to K means, FCM and ECM.

Acknowledgements

The authors want to thank anonymous reviewers for their remarks which helped us to improve the quality of this paper. This work has been partially supported by the China Natural Science Foundation (No. 61075029) and PhD Thesis Innovation Fund from Northwestern Polytechnical University (No. cx201015).

References

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New-York.

Dave, R.N., 1991. Clustering relational data containing noise and outliers. Pattern Recognition Lett. 12, 657–664.

Davies, D., Bouldin, D., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intell. 1 (2), 224–227.

Denœux, T., Masson, M.-H., 2003. Clustering of proximity data using belief functions. In: Bouchon-Meunier, B., Foulloy, L., Yager, R.R. (Eds.), Intelligent Systems for Information Processing from Representation to Application. Elsevier, Amsterdam, pp. 291–302.

Denœux, T., Masson, M.-H., 2004. EVCLUS: EVidential CLUStering of proximity data. IEEE Trans. Systems Man Cybernet. Part B 34 (1), 95–109.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugenic 7, 179–188, <http://www.en.wikipedia.org/wiki/Iris_flower_data_set>.

Krishnapuram, R., Keller, J., 1996. The possibilistic c-means algorithm: insights and recommendations. IEEE Trans. Fuzzy Systems 4 (3), 385–393.

Lloyd, S.P., 1982. Least squares quantization in PCM. IEEE. Trans. Inform. Theory 28 (2), 129–137.

Masson, M.-H., Denœux, T., 2004. Clustering interval-valued data using belief functions. Pattern Recognition Lett. 25 (2), 163–171.

Masson, M.-H., Denœux, T., 2008. ECM: An evidential version of the fuzzy c-means algorithm. Pattern Recognition 41 (4), 1384–1397.

Menard, M., Demko, C., Loonis, P., 2000. The fuzzy c+2-means: Solving the ambiguity rejection in clustering. Pattern Recognition 33, 1219–1237.

Shafer, G., 1976. A Mathematical Theory of Evidence. Princeton Univ. Press.

Smarandache, F., Dezert, J. (Eds.), Advances and Applications of DSmT for Information Fusion, American Research Press, Rehoboth, vol. 1–3, 2004–2009, <http://www.fs.gallup.unm.edu//DSmT.htm>

Smets, P., 1990. The combination of evidence in the transferable belief model. IEEE Trans. Pattern Anal. Machine Intell. 12 (5), 447–458.

Smets, P., 2005. Decision making in the TBM: the necessity of the pignistic transformation. Internat. J. Approx. Reason. 38.

Smets, P., Kennes, R., 1994. The transferable belief model. Artificial Intelligence 66, 191–243.

Windham, M.P., 1985. Numerical classification of proximity data with assignment measure. J. Classif. 2, 157–172.