BPEC: Belief-Peaks Evidential Clustering

Zhi-gang Su and Thierry Denoeux

***Abstract*—This paper introduces a new evidential clustering method based on the notion of "belief peaks" in the framework of belief functions. The basic idea is that all data objects in the neighborhood of each sample provide pieces of evidence that induce belief on the possibility of such sample to become a cluster center. A sample having higher belief than its neighbors and located far away from other local maxima is then characterized as cluster center. Finally, a credal partition is created by minimizing an objective function with the fixed cluster centers. An adaptive distance metric is used to fit for unknown shapes of data structures. We show that the proposed evidential clustering procedure has very good performance with an ability to reveal the data structure in the form of a credal partition, from which hard, fuzzy, possibilistic and rough partitions can be derived. Simulations on synthetic and real-world datasets validate our conclusions.**

***Index Terms*—Dempster-Shafer theory, belief functions, unsupervised learning, soft clustering, density peaks clustering.**

I. Introduction

CLUSTERING is one of the most important tasks in data mining and machine learning. It aims to find groups or clusters of objects that are similar to one another but dissimilar to objects in any other clusters. With different philosophies, distinct clustering techniques have been derived, for example, see [1]–[4] and the literature therein. Among them, *partitional clustering* has attracted a lot of attention in artificial intelligence communities.

Classical *hard partitional clustering* intends to assign each object unambiguously to one cluster with full certainty. Recently, Rodriguez and Laio proposed such a hard partitional clustering algorithm by fast search and find of density peaks, called *density peaks clustering* (DPC) [5]. In the DPC, a cluster center is defined as an object surrounded by neighbors with lower local densities and far away from any other object with higher local density. In order to detect all the cluster centers, density ρ_i

$$\rho_i = \sum\nolimits_{j \neq i} \chi(d_{ij}, d_c) \tag{1}$$

is first computed at each data object o_i according to distance d_{ij} (between objects o_i and $o_j, i = 1, 2, \cdots, n, j \neq i$) using a cutoff or Gaussian kernel $\chi(\cdot, \cdot)$ with cutoff distance d_c. Next, for each object, the distance δ_i separating it from its nearest object with a higher density is computed as

$$\delta_i = \begin{cases} \max_{1 \leq j \leq n}\{d_{ij}\}, & \text{if} \quad i = \arg\max_j\{\rho_j\}, \\ \min_{j:\rho_j > \rho_i}\{d_{ij}\}, & \text{otherwise.} \end{cases} \tag{2}$$

This work is supported in part by the National Natural Science Foundation of China under Grants 51876035, 51676034 and 51476028, and fully by the Key Project of Yunnan Power Grid Co. Ltd. under Grant YNYJ2016043.

Z.-G Su is with the School of Energy and Environment, Southeast University, Nanjing, Jiangsu 210096, China (e-mail: zhigangsu@seu.edu.cn).

T. Denoeux is with Sorbonne Universités, Université de Technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France (e-mail: thierry.denoeux@hds.utc.fr).

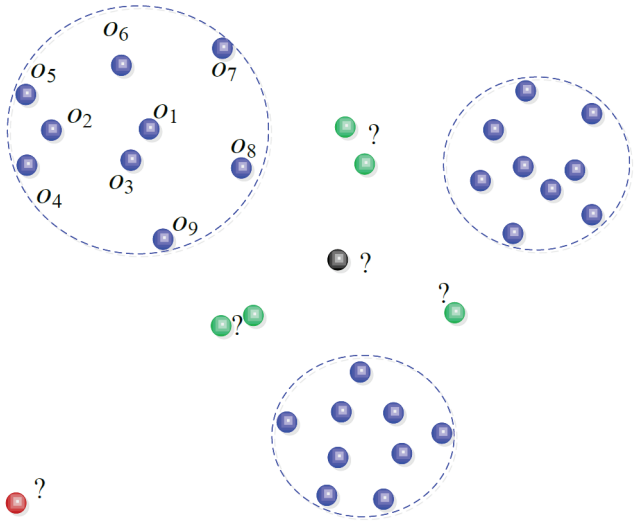

Fig. 1. Illustration of ambiguity and uncertainty in clustering.

By drawing a decision graph with ρ and δ as $x-$ and $y-$axes, respectively, cluster centers are then defined as the data objects that have both high density and large distance. At last, each of the remaining data objects is assigned heuristically to the same cluster as its nearest neighbor with higher density. One merit of the DPC algorithm is its ability to detect cluster centers without requiring to fix the number of clusters as priori. Therefore, a lot of interesting work on DPC has emerged. See, for example, [6]–[10].

However, the DPC and its variants compute hard partitions: they do not allow ambiguity, uncertainty or doubt (rather than noise) in the assignment of objects to clusters. As illustrated in Fig. 1, the objects between or among different clusters should be considered as ambiguous and/or uncertain. In contrat, *evidential clustering* [2], [11]–[16] allows us to describes ambiguity and uncertainty in the membership of objects to clusters using Dempster-Shafer mass functions [17]. Roughly speaking, a mass function can be seen as a collection of sets with corresponding masses. A collection of such mass functions for n objects is called *credal partition*.

Furthermore, in DPC algorithms, each object in the neighborhood of a sample provides just a numerical measure (i.e., a cutoff or Gaussian kernel function value of the distance between the object and the sample) supporting such sample to become a cluster center. As a matter of fact, an object in the neighborhood of a sample could provide more useful information on the possibility of such sample to become a cluster center. With this in mind and the theoretic viewpoint of belief functions, we may describe the support degree at each object by a mass function. The belief degree (associated to a mass function) at each data object can then be viewed as an extension of the (local) density in DPC algorithms. As we

will see, the cluster centers selected according to belief peaks are usually different from, and more appropriate than those obtained by density peaks. For instance, in Fig. 1, object o_2 can be selected as a cluster center according to density peaks, whereas the object o_1 is preferred by belief peaks, as will be discussed in Section III-A.

Motivating by the above considerations, this paper intends to propose a new evidential clustering method based on finding belief (rather than density) peaks as well as a credal partition in the theoretic framework of belief functions [17]–[19]. More precisely, all data objects in the neighborhood of each sample provide pieces of evidence on the possibility of such sample to become a cluster center. Then, by combing these pieces of evidence, a sample having higher belief than its neighbors and located far away from other local maxima will be characterized as a cluster center. Once all the cluster centers have been fixed, a credal partition will finally be created by minimizing an objective function, using an adaptive distance metric to describe non-spherical clusters. In this paper, we call our method *Belief-Peaks Evidential Clustering* (BPEC). The philosophy of BPEC is distinct from that of the DPC in several respects:

- BPEC selects cluster centers from the viewpoint of *information fusion* in the theoretic framework of belief functions, considering more useful information on the possibility of a data object to become a cluster center.
- BPEC creates a credal partition allowing ambiguity and uncertainty in the assignment of data objects, by solving a constrained optimization problem (with fixed cluster centers) as an alternative to heuristic assignment.
- The credal representation in the BPEC provides us a flexible way to reveal the data structure and, in particular, it can produce hard, fuzzy [20], possibilistic [21] and rough [22], [23] partitions.

As will be shown in Section IV, the BPEC procedure has good performances and outperforms the standard DPC algorithm as well as some other evidential clustering algorithms in most cases.

The rest of this paper is organized as follows. The theory of belief functions and the notion of credal partition are first briefly recalled in Section II. The BPEC method is then introduced in Section III. In Section IV, we conduct some experiments to study the performances of BPEC using some synthetic and real-world datasets. The last section concludes the paper.

II. Preliminaries

A. Background on belief functions

In this subsection, we briefly recall some basic notions of the theory of belief functions [17]–[19], [24]–[26] needed in the rest of the paper. Given a *frame of discernment* $\Omega = \{\omega_1, \omega_2, \cdots, \omega_c\}$, a mass function is defined as a mapping from 2^Ω to $[0, 1]$ such that

$$\sum\nolimits_{A\subseteq\Omega} m^\Omega(A) = 1. \tag{3}$$

The subsets A of Ω such that $m^\Omega(A) > 0$ are called the *focal sets* of m^Ω. A mass function is said to be

- *Bayesian* if it only has singletons (i.e., $|A| = 1$) as focal sets, and *unnormalized Bayesian* if it has either singletons or the empty set ($\emptyset$) as focal sets;
- *Consonant* if its focal sets are nested;
- *Logical* if it has only one nonempty focal set;
- *Non-dogmatic* if it has Ω as one focal set; in particular, the vacuous mass function, verifying $m^\Omega(\Omega) = 1$, corresponds to total ignorance;
- *Unnormalized* if it has the empty set as one focal set, and *normalized* otherwise.

There are other equivalent representations of a mass function such as the *belief* and *plausibility* functions defined, respectively, as

$$Bel^\Omega(A) = \sum_{\emptyset\neq B\subseteq A} m^\Omega(B), \tag{4}$$

$$Pl^\Omega(A) = \sum_{A\cap B\neq\emptyset} m^\Omega(B), \tag{5}$$

for all $A \subseteq \Omega$. The function $pl^\Omega : \Omega \to [0, 1]$ such that $pl^\Omega(\omega) = Pl^\Omega(\{\omega\})$ is called the *contour function* associated to m^Ω. If m^Ω is Bayesian, we have $pl^\Omega(\omega) = m^\Omega(\{\omega\})$ for all $\omega \in \Omega$. In this case, pl^Ω is a probability distribution.

The combination of mass functions plays a critical role in the theory of belief functions. Let m_1 and m_2 be two mass functions. The conjunctive combination of m_1 and m_2 yields the unnormalized mass function

$$m^\Omega_{1\cap 2}(A) = \sum_{B\cap C=A} m^\Omega_1(B) m^\Omega_2(C),\ \forall A \subseteq \Omega. \tag{6}$$

If necessary, the normality condition $m^\Omega(\emptyset) = 0$ may be recovered by dividing each mass $m^\Omega_{1\cap 2}(A)$ by $1 - m^\Omega_{1\cap 2}(\emptyset)$. The resulting operation is noted $\oplus$ and is called *Dempster's rule of combination*:

$$m^\Omega_{1\oplus 2}(A) = \frac{m^\Omega_{1\cap 2}(A)}{1 - m^\Omega_{1\cap 2}(\emptyset)},\ \emptyset \neq A \subseteq \Omega. \tag{7}$$

Both rules are commutative, associative and admit the vacuous mass function as a unique neutral element.

B. Credal partitions

Suppose that we have a set $\mathcal{O} = \{o_1, o_2, \cdots, o_n\}$ of n objects. Let $\Omega = \{\omega_1, \omega_2, \cdots, \omega_c\}$ be the set of c clusters. If we know for sure which cluster each object belongs to, we have a hard partition of the dataset $\mathcal{O}$. More precisely, a hard partition may be represented by binary variables u_{ik} such that $u_{ik} = 1$ if object o_i belongs to cluster ω_k, and $u_{ik} = 0$ otherwise.

When objects cannot be assigned to clusters with certainty, one can represent ambiguous and uncertain cluster memberships by mass functions $m^\Omega_i, i = 1, 2, \cdots, n$. Each mass $m^\Omega_i(A)$ is interpreted as a degree of support attached to the proposition "the true cluster of object o_i is in A", and to no more specific proposition. The $n-$tuple $\mathcal{M}^\Omega = (m^\Omega_1, m^\Omega_2, \cdots, m^\Omega_n)$ is called a credal partition [2], [13].

Example 1: The $4-$tuple $\mathcal{M}^\Omega = (m^\Omega_1, m^\Omega_2, m^\Omega_3, m^\Omega_4)$ in Table I is an example of a credal partition. We can see that objects o_1 and o_3 likely belong to ω_1 and ω_2, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2018.2869125, IEEE Transactions on Fuzzy Systems

TABLE I
A CREDAL PARTITION ON $\Omega = \{\omega_1, \omega_2\}$

	$\emptyset$	$\{\omega_1\}$	$\{\omega_2\}$	Ω
m_1^Ω	0	0.7	0.2	0.1
m_2^Ω	0	0	0	1.0
m_3^Ω	0	0.2	0.8	0
m_4^Ω	1.0	0	0	0

In contrast, objects o_2 and o_4 correspond to two different situations of maximum uncertainty. Object o_2 has a full mass assigned to Ω: this reflects total ambiguity in the class membership of this object, which means that it might belong to ω_1 as well as to ω_2. For objet o_4, it has the largest mass assigned to the empty set, indicating that this object does not seem to belong to any of the two clusters and is an *outlier*. □

The notion of credal partition boils down to several other usual clustering structures when the mass functions have some special types [2].

- **Hard partition:** We have a hard partition with $u_{ik} = 1$ if all mass functions m_i^Ω are certain, i.e., $m_i^\Omega(\{\omega_k\}) = 1$ for some k, and $u_{ik} = 0$ otherwise.
- **Fuzzy partition:** We have a fuzzy partition with $u_{ik} = m_i^\Omega(\{\omega_k\})$ if all the mass functions m_i^Ω are Bayesian. In particular, a fuzzy partition with a noise cluster may be obtained if all the mass functions m_i^Ω are unnormalized Bayesian, i.e., if $u_{ik} = m_i^\Omega(\{\omega_k\})$ and $u_{i*} = m_{i\emptyset}^\Omega = m_i^\Omega(\emptyset) = 1 - \sum_{k=1}^c u_{ik}$.
- **Possibilistic partition:** if masses m_i^Ω are consonant, the corresponding plausibility function is formally a possibility measure, and the credal partition can be seen as a possibilistic partition, $u_{ik} = pl_i^\Omega(\omega_k)$ being interpreted as the possibility that object o_i belongs to cluster ω_k.
- **Rough partition:** if masses m_i^Ω are logical, i.e., we have $m_i^\Omega(A_i) = 1$ for some $\emptyset \neq A_i \subseteq \Omega$, we can then define the lower and upper approximations of cluster ω_k as

$$\omega_k^L = \{o_i | A_i = \{\omega_k\}\}, \quad \omega_k^U = \{o_i | \{\omega_k\} \subseteq A_i\}. \quad (8)$$

  The memberships to the lower and upper approximations are then, respectively, $\underline{u}_{ik} = Bel_i^\Omega(\{\omega_k\})$ and $\bar{u}_{ik} = Pl_i^\Omega(\{\omega_k\})$.

In general the credal partition obtained by an evidential clustering algorithm does not belong to any of the previous specific types, but it can be transformed into a simpler representation. A fuzzy partition can be obtained by transforming each mass function m_i^Ω into a probability distribution p_i using the *plausibility-probability* transformation defined by

$$p_i(\omega_k) = \frac{pl_i^\Omega(\omega_k)}{\sum_{l=1}^c pl_i^\Omega(\omega_l)}, \ k = 1, 2, \cdots, c. \quad (9)$$

By selecting the cluster with maximum probability for each object, we get a hard partition. To obtain a fuzzy partition with a noise cluster, the degree of membership of each object o_i to cluster ω_k can be defined as $u_{ik} = (1 - m_{i\emptyset}^\Omega) p_i(\omega_k)$ and the degree of membership to the noise cluster as $u_{i*} = m_{i\emptyset}^\Omega$. To obtain rough partition, for instance, the following *interval dominance decision rule* [13] can be used to select the set A_i of clusters whose plausibility exceeds the degree of belief of any other clusters,

$$A_i = \{\omega \in \Omega | \forall \omega' \in \Omega, pl_i^\Omega(\omega) \geq Bel_i^\Omega(\{\omega'\})\}, \quad (10)$$

where pl_i^Ω and Bel_i^Ω are the normalized contour and belief functions associated to mass function m_i^Ω.

III. PROPOSED METHOD: BPEC

Similar to DPC algorithms, the BPEC algorithm also consists of two parts: definition of cluster centers and assignment of the remaining data objects. These two parts will be discussed in Sections III-A and III-B, respectively. In Section III-C, the time complexity of BPEC will be analyzed and a variant with a limited number of informative composite clusters will be introduced. The tuning of parameters will be addressed in Section III-D.

A. Belief peaks

For a given set $\mathcal{O}$ of n data objects, a new frame of discernment $\mathcal{C} = \{C, \neg C\}$ is defined to discern whether an object is a cluster center (C) or not ($\neg C$). The basic idea to detect cluster centers can be summarized as follows. Let $\mathcal{N}_K(o_i)$ denote the set of the K nearest neighbors (KNN) of object o_i. Each neighbor o_j in $\mathcal{N}_K(o_i)$ provides a piece of evidence about object o_i being a cluster center. This piece of evidence can be represented by a mass function $m_{ij}^\mathcal{C}$. By combining these mass functions using Dempster's rule (6)-(7), a normalized mass function $m_i^\mathcal{C}$ can be obtained as well as its associated belief function $Bel_i^\mathcal{C}$. An object having higher degree of belief $Bel_i^\mathcal{C}(\{C\})$ than its neighbors will be characterized as a cluster center if it is also at a relatively large distance from other objects with higher degrees of belief.

Each neighbor o_j in $\mathcal{N}_K(o_i)$ supports the assumption that o_i is a cluster center if the distance d_{ij} between the two objects is small. If this distance is large, the evidence of object o_j is inconclusive. Mass function $m_{ij}^\mathcal{C}$ can, thus, be defined as

$$m_{ij}^\mathcal{C}(A) = \begin{cases} \phi(d_{ij}^2), & A = \{C\}, \\ 1 - \phi(d_{ij}^2), & A = \mathcal{C}, \end{cases} \quad (11)$$

where $\phi(\cdot)$ is a decreasing function verifying

$$\phi(0) = \alpha_0 \text{ and } \lim_{d_{ij}^2 \to \infty} \phi(d_{ij}^2) = 0 \quad (12)$$

with a constant α_0 such that $0 < \alpha_0 \leq 1$. A popular choice for $\phi(\cdot)$ is [27], [28]

$$\phi(d_{ij}^2) = \alpha_0 \exp(-\gamma_i^2 d_{ij}^2), \quad (13)$$

where γ_i is a positive parameter associated to object o_i.

Using Dempster's rule, the final normal mass function $m_i^\mathcal{C}$ for object o_i can be calculated as

$$m_i^\mathcal{C} = \bigoplus_{j \in \mathcal{N}_K(o_i)} m_{ij}^\mathcal{C}. \quad (14)$$

According to (14), we have the following proposition used to calculate belief $Bel_i^\mathcal{C}(\{C\})$ at data object o_i.

Proposition 1: The degree of belief that object o_i is a cluster center is

$$Bel_i^{\mathcal{C}}(\{C\}) = 1 - \prod_{j \in \mathcal{N}_K(o_i)} \left[1 - \phi(d_{ij}^2)\right]. \quad (15)$$

Proof: The combined mass function has two focal sets: $\mathcal{C}$ and $\{C\}$. The mass on $\mathcal{C}$ is the product of the $1 - \phi(d_{ij})$, so the mass on $\{C\}$ is 1 minus this product. Hence, we have

$$m_i^{\mathcal{C}}(\mathcal{C}) = \prod_{j \in \mathcal{N}_K(o_i)} \left[1 - \phi(d_{ij}^2)\right] \quad (16)$$

and

$$m_i^{\mathcal{C}}(\{C\}) = 1 - \prod_{j \in \mathcal{N}_K(o_i)} \left[1 - \phi(d_{ij}^2)\right]. \quad (17)$$

Now, $Bel_i^{\mathcal{C}}(\{C\}) = m_i^{\mathcal{C}}(\{C\})$, which completes the proof. ■

Proposition 1 provides us with the final belief on the possibility of an object to become a cluster center. According to the basic idea of BPEC, to be a cluster center an object should not only have high degree of belief, but it should also be located far away from other objects with high degrees of belief. Hence, a metric should be defined in order to measure distances among objects, as that done in DPC. Here, we redefine the *delta* metric δ_i in (2), as follow

$$\delta_i = \min_{\{j: Bel_j^{\mathcal{C}}(\{C\}) > Bel_i^{\mathcal{C}}(\{C\})\}} d_{ij} \quad (18)$$

for objects that do not have the highest degree of belief and $\delta_i = \max_{1 \le j \le n}\{d_{ij}\}$ for the object with the highest degree of belief.

Now, we can construct a $\delta - Bel^{\mathcal{C}}(\{C\})$ ($\delta - Bel$ for short) decision graph by plotting $Bel^{\mathcal{C}}(\{C\})$ versus δ. The objects with higher $Bel_i^{\mathcal{C}}(\{C\})$ and larger δ_i are identified as cluster centers. As illustrated in Fig. 2, cluster centers usually appear in the upper right corner of the decision graph. For given lower bounds δ_{min} and Bel_{min}, the objects such that $\delta_i > \delta_{min}$ and $Bel_i^{\mathcal{C}}(\{C\}) > Bel_{min}$ will be selected as centers. Meanwhile, data objects with small degrees of belief and large deltas can be detected as outliers from the decision graph.

Remark 1: There exist some relationships between the belief ($Bel_i^{\mathcal{C}}(\{C\})$) and the density ($\rho_i$) in DPC algorithms. Four typical densities ρ_i are briefly recalled as follows. With Gaussian kernel, i.e., $\chi(d_{ij}, d_c) = \phi(d_{ij}^2)$ with $\alpha_0 = 1$ and $\gamma_i = 1/d_c$, density (1) can be rewritten as $\rho_i := \sum_{j=1}^n \phi(d_{ij}^2)$. To reduce the influence of data objects far away, the density can be locally defined as $\rho_i := \sum_{j \in \mathcal{N}_K(o_i)} \phi(d_{ij}^2)$, e.g., see the adaptive DPC (ADPC-KNN) [9] in which $\gamma_i = 1/(\mu^K + \sigma^K)$ such that $\mu^K = \frac{1}{n}\sum_{i=1}^n \nu_i^K$, $\sigma^K = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (\nu_i^K - \mu^K)^2}$ and $\nu_i^K = \max_{j \in \mathcal{N}_K(o_i)}(d_{ij})$. In [10], Xie et al. proposed a fuzzy weighted KNN-DPC (FKNN-DPC) in which the local density is defined as $\rho_i := \sum_{j \in \mathcal{N}_K(o_i)} \phi(d_{ij})$ with $\alpha_0 = 1$, $\gamma_i = 1$ and non-squared Euclidean distance. In a different way, Du et al. [7] proposed a DPC-KNN algorithm in which the local density was defined as $\rho_i := \prod_{j \in \mathcal{N}_K(o_i)} \phi(d_{ij}^2)$ with $\alpha_0 = 1$ and $\gamma_i = 1/\sqrt{nf}$, where f is a constant percentage. By expanding (15), we get

$$Bel_i^{\mathcal{C}}(\{C\}) = (-1)^{K+1} \prod\nolimits_{j \in \mathcal{N}_K(o_i)} \phi(d_{ij}^2) + \sum\nolimits_{j \in \mathcal{N}_K(o_i)} \phi(d_{ij}^2) + l(\phi(d_{ij}^2)) \quad (19)$$

with a polynomial function $l(\cdot)$. It is evident that the belief $Bel_i^{\mathcal{C}}(\{C\})$ can be viewed as an extension of densities defined in [5] with $K = n$, [9] and [7]. If one uses a non-squared distance, then $Bel_i^{\mathcal{C}}(\{C\})$ is an extension of the local density in [10]. The distribution of cluster centers via the BPEC is usually different from and arguably more appropriate than those obtained by these DPC algorithms. To show this, a numerical example is presented as follows, and another synthetic example will be presented later in Example 4. □

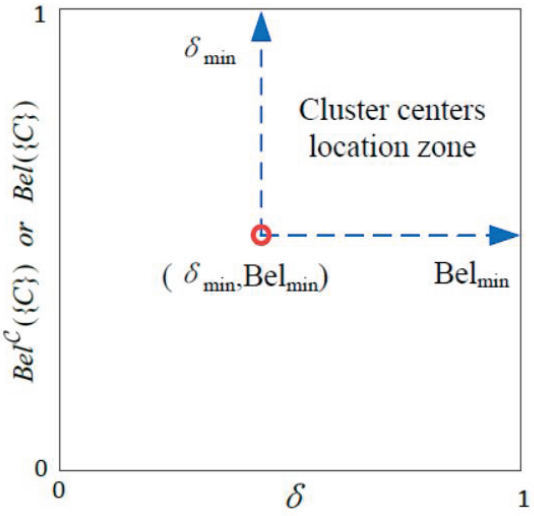

Fig. 2. An example of decision graph $\delta - Bel$

Example 2: Given data objects o_i, $i = 1, 2, \cdots, 9$ as illustrated in Fig. 1. Suppose $\phi(d_{12}^2) = 0.3$, $\phi(d_{13}^2) = 0.9$, $\phi(d_{16}^2) = 0.45$, $\phi(d_{23}^2) = 0.35$, $\phi(d_{24}^2) = \phi(d_{25}^2) = 0.7$, $\phi(d_{36}^2) = 0.25$, $\phi(d_{45}^2) = 0.5$, and all $\phi(\cdot)$ between any other two objects are smaller than 0.2. Furthermore, we assume $\phi(d_{ij}^2) = \phi(d_{ji}^2)$, $i, j = 1, 2, \cdots, 9$. For simplicity, this example considers $\rho_i = \sum_{j \in \mathcal{N}_K(o_i)} \phi(d_{ij}^2)$ with $K = 3$. We have $\rho_1 = 0.45 + 0.9 + 0.3 = 1.65$, $\rho_2 = 0.7 \times 2 + 0.35 = 1.75$, $\rho_3 = 0.9 + 0.35 + 0.25 = 1.5$, $\rho_4, \rho_5 < 0.5 + 0.7 + 0.2 = 1.4$, $\rho_6 < 0.45 + 0.25 + 0.2 = 0.9$, $\rho_i < 0.2 \times 3 = 0.6$, $i = 7, 8, 9$. According to (15), we can obtain $Bel_1^{\mathcal{C}}(\{C\}) = 1 - 0.1 \times 0.55 \times 0.7 = 0.9615$, $Bel_2^{\mathcal{C}}(\{C\}) = 1 - 0.3^2 \times 0.65 = 0.9415$, $Bel_3^{\mathcal{C}}(\{C\}) = 1 - 0.1 \times 0.65 \times 0.75 = 0.95125$, $Bel_4^{\mathcal{C}}(\{C\}), Bel_5^{\mathcal{C}}(\{C\}) < 1 - 0.3 \times 0.5 \times 0.8 = 0.88$, $Bel_6^{\mathcal{C}}(\{C\}) < 1 - 0.55 \times 0.75 \times 0.8 = 0.67$, $Bel_i^{\mathcal{C}}(\{C\}) < 1 - 0.8^3 = 0.488$, $i = 7, 8, 9$. Hence, $\rho_2 = \max_{1 \le i \le 9}\{\rho_i\}$ while $Bel_1^{\mathcal{C}}(\{C\}) = \max_{1 \le i \le 9}\{Bel_i^{\mathcal{C}}(\{C\})\}$. Consequently, object o_2 is selected as a cluster center by the density peaks method, whereas o_1, located closer to the center of gravity of these nine data objects, is preferred by the belief peaks method. □

Remark 2: In some cases, some prior knowledge may be obtained in the form of mass functions $m_{i0}^{\mathcal{C}}, i = 1, 2, \ldots, n$,

representing the initial possibility of each data object to be selected as a cluster center. In this case, the final masses $m_i^{\mathcal{C}}$ in (14) can be recalculated as [1]

$$m_i^{\mathcal{C}} = m_{i0}^{\mathcal{C}} \bigoplus \left(\bigoplus_{j \in \mathcal{N}_K(o_i)} m_{ij}^{\mathcal{C}} \right). \tag{20}$$

However, it is not clear how the prior mass functions $m_{i0}^{\mathcal{C}}$ could be obtained in practice. We leave this problem for further research. □

B. Construction of the credal partition

In this section, we address the problem of deriving a credal partition $\mathcal{M}^\Omega = (m_1^\Omega, m_2^\Omega, \ldots, m_n^\Omega)' \in \mathbb{R}^{n \times 2^c}$ for the set of objects $\mathcal{O}$ locating at $(x_1, x_2, \ldots, x_n)' \in \mathbb{R}^{n \times p}$, by minimizing an objective function with the fixed cluster centers $v_k \in \mathbb{R}^p, k = 1, 2, \ldots, c$, found in the previous section.

Deriving a credal partition from $\mathcal{O}$ implies determining the quantities $m_{ij}^\Omega = m_i^\Omega(A_j), A_j \subseteq \Omega$ for each object o_i in such a way that m_{ij}^Ω is high (respectively, low) when the distance D_{ij} between o_i and the focal set A_j is small (respectively, large). Suppose each cluster ω_k is represented by a center v_k. As suggested in the *Evidential C-Means* algorithm (ECM) [15], the barycenter $\bar{v}_j$ of the centers associated to the clusters composing nonempty set A_j can be defined as

$$\bar{v}_j = \frac{1}{|A_j|} \sum_{k=1}^{c} s_{kj} v_k, \tag{21}$$

where $s_{kj} = 1$ if $\omega_k \in A_j$ and $s_{kj} = 0$ otherwise. Let S_k be the $p \times p$ symmetric positive definite matrix associated to cluster ω_k inducing a norm $\|x\|_{S_k}^2 = x' S_k x$. Similar to (21), we can define the matrix $\bar{S}_j$ associated to the nonempty subset (i.e., the composite cluster) A_j. The distance between object o_i and composite cluster A_j is then

$$\begin{aligned} D_{ij}^2 &= \|x_i - \bar{v}_j\|_{\bar{S}_j}^2 \\ &= (x_i - \bar{v}_j)' |A_j|^{-1} \sum_{k=1}^{c} s_{kj} S_k (x_i - \bar{v}_j). \end{aligned} \tag{22}$$

Then, a credal partition $\mathcal{M}^\Omega$ can be derived by minimizing the following objective function:

$$\begin{aligned} &\mathcal{J}_{BPEC}(\mathcal{M}^\Omega, S_1, \cdots, S_c) \\ &\quad = \sum_{i=1}^{n} \sum_{j: A_j \neq \emptyset} |A_j|^\alpha (m_{ij}^\Omega)^\beta D_{ij}^2 + \sum_{i=1}^{n} \Delta^2 (m_{i\emptyset}^\Omega)^\beta, \end{aligned} \tag{23}$$

subject to

$$\begin{cases} \sum_{j: A_j \neq \emptyset} m_{ij}^\Omega + m_{i\emptyset}^\Omega = 1, & i = 1, 2, \cdots, n, \\ \det(S_k) = 1, & k = 1, 2, \cdots, c, \end{cases} \tag{24}$$

where constants α, β and Δ have the same meaning as those in the ECM algorithm, and $\det(\cdot)$ denotes the determinant of a matrix. Note that the empty focal set $\emptyset$ is treated separately from other nonempty focal sets by using a constant Δ.

[1]This was suggested by an anonymous referee.

To minimize $\mathcal{J}_{BPEC}$, we apply the alternate optimization algorithm as suggested in the constrained ECM [11] and get

$$m_{ij}^\Omega = \frac{|A_j|^{-\alpha/(\beta-1)} D_{ij}^{-2/(\beta-1)}}{\sum_{l: A_l \neq \emptyset} |A_l|^{-\alpha/(\beta-1)} D_{il}^{-2/(\beta-1)} + \Delta^{-2/(\beta-1)}}, \tag{25a}$$

$$m_{i\emptyset}^\Omega = 1 - \sum\nolimits_{j: A_j \neq \emptyset} m_{ij}^\Omega, \tag{25b}$$

for $i = 1, 2, \cdots, n, j : \emptyset \neq A_j \subseteq \Omega$, and

$$S_k = \det(\Sigma_k)^{1/p} \Sigma_k^{-1}, \quad k = 1, 2, \cdots, c, \tag{26}$$

with

$$\Sigma_k = \sum_{i=1}^{n} \sum_{j: A_j \neq \emptyset} |A_j|^{\alpha-1} (m_{ij}^\Omega)^\beta s_{kj} (x_i - \bar{v}_j)(x_i - \bar{v}_j)'.$$

For completeness, the calculations of (25) and (26) are presented in the Appendix A.

The BPEC algorithm can be summarized in Algorithm 1.

Algorithm 1: Belief-peaks evidential clustering

Input: $K, \alpha_0, q, \alpha, \beta$, termination threshold ε, Δ, δ_{min}, Bel_{min}, data of objects $x_i \in \mathbb{R}^p, i = 1, 2, \cdots, n$

Calculate degrees of belief ($Bel_i^{\mathcal{C}}(\{\mathrm{C}\})$) for all objects using (15)
Calculate delta's (δ_i) for all objects according to (18)
Draw the decision graph $\delta - Bel$, and determine lower bounds δ_{min} and Bel_{min}
Select cluster centers $v_k, k = 1, 2, \cdots c$ in the decision graph and determine Δ
$t \leftarrow 0, S_k(0) = I, \mathcal{M}^\Omega(0) = 0$
repeat
  $t \leftarrow t + 1$; % t is the number of iterations
  Calculate $\mathcal{M}^\Omega(t)$ according to (21), (22) and (25) using $S_k(t-1)$;
  Update matrix $S_k(t)$ according to (26)
until $\|\mathcal{M}^\Omega(t) - \mathcal{M}^\Omega(t-1)\| < \varepsilon$;

Output: cluster centers v_k and credal partition $\mathcal{M}^\Omega$

Remark 3: When selecting cluster centers, BPEC considers the local geometric information of each data object, whereas the ECM algorithm does not. This characteristic may result in more appropriate cluster centers and, thus, a more reasonable credal partition, as shown later in Example 3. Furthermore, when just applying $\delta - Bel$ instead of $\rho - \delta$ in DPC [5], namely, selecting cluster centers according to belief peaks and assigning each of the remaining objects to the same cluster as its nearest neighbor with higher belief, we can intuitively induce a *belief version* of DPC, called *Belief peaks clustering* (BPC), which is summarized in Algorithm 2 and is presented in Appendix B. The BPC will be implemented under the same initial conditions as those of the BPEC algorithm. □

Remark 4: The barycenters (21) as defined in ECM may lead to uninformative composite clusters in some cases, as remarked in [12]. To reduce the time complexity of BPEC, an informative BPEC algorithm with a limited number of composite clusters will be discussed in Section III-C. □

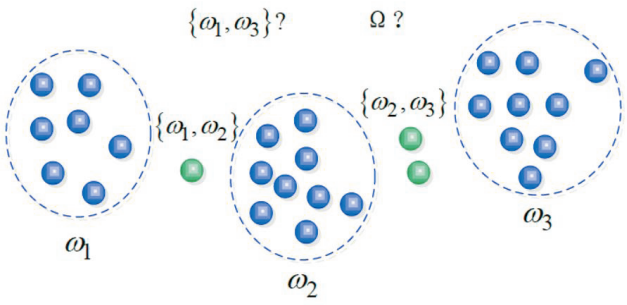

Fig. 3. Illustration of less informative composite clusters in a credal partition.

C. Time complexity analysis and informative BPEC

Using the notations of Algorithm 1, the time complexity of BPEC can be analyzed as follows.

First, the computation cost to obtain the belief peaks in Step 1 is $O(p^2n(n-1)/2) + O(2^2Kn)$. Then, the sorting and assignment processes in Step 2 can be performed in $O(n\log n) + O(n)$ operations. Finally, to derive a credal partition in Steps $6-10$, we need to perform $O(2^ctn)+O(2^ct)$ operations. The total time complexity of BPEC is, thus, $O(p^2n(n-1)/2 + 2^ct(n+1) + n(\log n + 2^2K + 1)) \sim O(p^2n^2 + 2^cnt)$.

The complexity of BPEC depends heavily on both the number n of objects and the number c of clusters. When c is large, we may have $2^c > n$, and the complexity of BPEC becomes prohibitive if the number c of clusters is too large. Moreover, some composite clusters (i.e., focal sets) in the credal partition may be less informative in some cases. For instance, in a two dimensional clustering task as shown in Fig. 3, the composite clusters $\{\omega_1, \omega_3\}$ and Ω are less informative than the other clusters.

Hence, it is interesting to remove less informative focal sets so as to reduce the time complexity of the BPEC algorithm. In [14], Denœux proposed to preserve the needed expressivity of the credal partition by considering as focal sets the empty set, the singletons, and optionally the whole frame of discernment, and then to add some informative pairs of clusters. As a result, the number of focal sets is much smaller than 2^c, in particular when the number c is large. In a similar way, the *informative BPEC* can be described as follows:

1) **Step 1:** Run the BPEC in the basic configuration with focal sets of cardinalities 0, 1, and optionally, c. An initial credal partition $\mathcal{M}_0^\Omega$ is then created. The similarity between each pair of clusters (ω_j, ω_l) is measured by

$$\mathcal{S}(j,l) = \sum\nolimits_{i=1}^{n} pl_{ij}^{\Omega} pl_{il}^{\Omega}, \tag{27}$$

where pl_{ij}^{Ω} and pl_{il}^{Ω} are, respectively, the normalized plausibilities that object o_i belongs to clusters ω_j and ω_l. We then can determine the set $\mathcal{P}_\mathcal{K}$ of pairs $\{\omega_j, \omega_l\}$ that are mutual $\mathcal{K}$ nearest neighbors according to similarity measure (Note that $\mathcal{K}$ should not be confused with K).
2) **Step 2:** Run the BPEC again starting from $\mathcal{M}_0^\Omega$, and adding as focal sets the pairs of clusters in $\mathcal{P}_\mathcal{K}$.

D. Tuning of parameters

To implement the BPEC algorithm, some parameters should be selected in advance, including K, α_0, γ_i, δ_{min}, Bel_{min}, α, β, Δ and ε. When implementing the informative BPEC, the number of mutual $\mathcal{K}$ nearest pairs of clusters also needs to be determined.

The constant α_0 in the decreasing function $\phi(\cdot)$ should be positive and not greater than 1. We suggest $\alpha_0 = 1/K$ in order to avoid "saturated beliefs" if the number of nearest neighbors K is large. (This issue will be discussed later in Remark 6).

Parameters γ_i and K together play critical roles on determining the distribution of objects in the decision graph. As shown in Remark 1, most DPC algorithms usually define a unique value of γ_i for all objects o_i, $i = 1, 2, \cdots, n$. It is interesting to allow "adaptive" γ_i (i.e., different cutoff distances) for different data objects. Here, we define γ_i as the inverse of a quantile of the distances between object o_i and its K nearest neighbors

$$\gamma_i = 1/\operatorname*{quantile}_{j\in\mathcal{N}_K(o_i)}(d_{ij}, q), \tag{28}$$

where q is a quantile number such that $0 \le q \le 1$. To fix γ_i in an automatic way for simplicity, we set $q = 0.9$ in this paper.

For KNN-based classification or clustering algorithms, there is no efficient rule to determine the optimal number K of neighbors automatically. For BPEC, a simple approach is to increase K until some objects (i.e., cluster centers and outliers) can be visually separated from the other objects. It will be seen that, for most datasets, a large value of K is preferable. More interestingly, the distribution of cluster centers in the decision graph is not too sensitive to the change of K provided K is large enough. More details will be given with experimental results in Section IV.

In contrast to K and q, it is easy to determine the lower bounds δ_{min} and Bel_{min}, because cluster centers can usually be located far away from other data objects. In fact, we will see that δ_{min} and Bel_{min} are not as crucial as they seem to be and there are many choices for δ_{min} and Bel_{min}. As recommended with most DPC algorithms, we suggest determining δ_{min} and Bel_{min} by visual inspection after drawing the decision graph. The lower bounds δ_{min} and Bel_{min} will be presented together with cluster centers in the decision graph for convenience.

To derive a credal partition, we should preset $\varepsilon, \alpha, \beta, \mathcal{K}$ and Δ. We use as default values $\varepsilon = 10^{-3}, \alpha = 1$ and $\beta = 2$, as in the ECM algorithm [15]. Furthermore, we just consider at most two mutual nearest pairs of clusters when the number of clusters is large (i.e., $\mathcal{K} = 1$ or 2). Finally, we set Δ^2 equal to a constant smaller than the minimal delta associated to outliers in the decision graph if outliers exist; otherwise, Δ^2 can be a constant larger than the maximum $\max_{1\le i\le n}\{\delta_i\}$.

Remark 5: As a matter of fact, tuning K and q together usually provides a more flexible way to distinguish cluster centers from other regular data objects. In practice, one can alternately increase and/or decrease K and q in their ranges until one gets an interpretable $\delta - Bel$ decision graph. □

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2018.2869125, IEEE Transactions on Fuzzy Systems

IV. Experimental results

This section consists of two parts. In Section IV-A, some numerical examples are used to illustrate the behavior of BPEC algorithm. In Section IV-B, we compare the performance of BPEC as to those of some other clustering algorithms.

During the simulations, all the attributes of data objects were normalized into $[0, 1]$ to make the results independent from the units in which these attributes are expressed, according to the following min-max rule:

$$x_{ij} \leftarrow \frac{x_{ij} - \min(x_{\cdot j})}{\max(x_{\cdot j}) - \min(x_{\cdot j})}, \tag{29}$$

where x_{ij} denotes the value of attribute j of object o_i, and $\min(x_{\cdot j})$ and $\max(x_{\cdot j})$ are, respectively, the minimum and maximum values of attribute j.

The *Adjust Rand Index* (ARI) [29] is a popular choice for a performance index and is suitable to measure the closeness of a hard partition to the truth. To perform comparisons among hard, fuzzy, rough and even credal partitions, Denœux proposed a credal version of ARI, called *Credal Rand Index* (CRI) [30]. In this paper, we use CRI as the performance index when comparing the closeness between two credal partitions. When comparing two hard partitions, CRI and ARI are equivalent. In this case, the criterion will be referred to as "ARI". We refer the reader to [29] and [30] for the precise definitions of ARI and CRI.

A. Illustrative examples

In the following three examples, we use the notation Bel in place of $Bel_i^{\mathcal{C}}(\{C\})$ for simplicity when there is no risk of confusion.

Example 3: In this example, we consider the famous *butterfly* dataset [12], [15] to illustrate the results of BPEC. The butterfly dataset is represented by circles in Fig. 4 (top). We set $K = 4,\ q = 0.9$ and $\Delta^2 = 0.1$. The $\delta - Bel$ decision graph can first be drawn, as shown in Fig. 5. In this graph, objects 3 and 9 have equally high degrees beliefs and are far away from other data objects. Hence, these two data objects can be considered as the centers of two clusters, i.e., $c = 2$. With the selected cluster centers, outlined by symbol " + " in Fig. 4 (top), a credal partition can then be created and illustrated in Fig. 4 (bottom), from which we can see that the credal partition is meaningful to reveal ambiguity and uncertainty. For instance, object o_6 could be assigned to any of the two detected clusters, whereas object o_{12} and o_{13} could be better identified as outliers. In contrast, o_6, o_7 and o_{13} were assigned to Ω by ECM, and instead, o_{13} was assigned to Ω by the Belief C-Means (BCM) algorithm [12]. More interestingly, BPEC can find the true locations of centers whereas the ECM and BCM cannot. □

Example 4: In this example, we consider the strongly overlapping *four-class* dataset [15] consisting of 100 of data objects in each class. We aim to study the influence of K, q and the locations of centers on the clustering performances. We considered $K \in \{20, 25, \cdots, 90\}$, $q \in \{0.1, 0.2, \cdots, 1.0\}$ and $\Delta^2 = 0.2$. In each case, we selected the cluster centers from the decision graph. The credal partition was then created

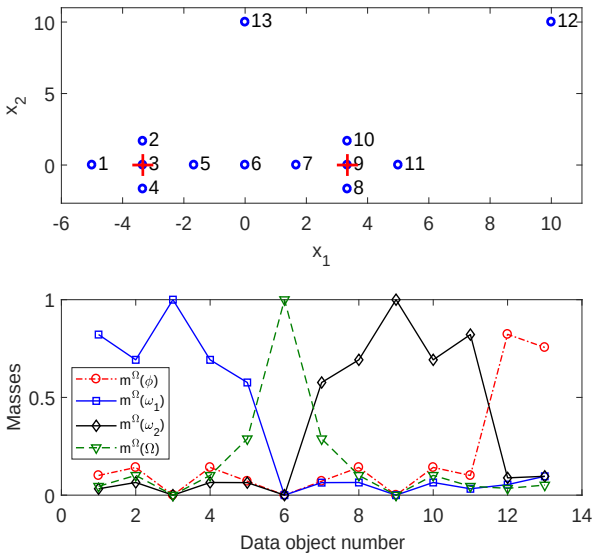

Fig. 4. *Butterfly* (top) and credal partition (bottom) via BPEC algorithm.

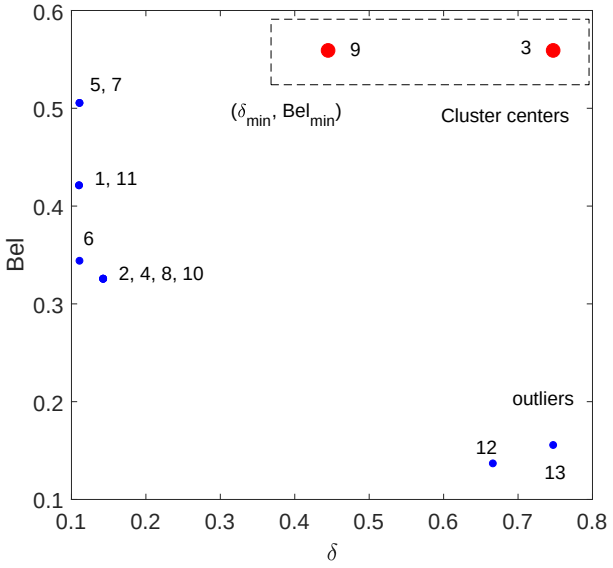

Fig. 5. $\delta - Bel$ decision graph for the butterfly dataset.

by BPEC. Fig. 6 shows the number of clusters and the ARI as functions of K and q. The ARI was computed after transforming the credal partition into a hard one using the maximum probability rule (9). From Fig. 6a, we can see that the true number of clusters can be easily identified in most cases, in particular when K and q take large values simultaneously. Correspondingly, Fig. 6b indicates that better performances (i.e., higher ARI) can be achieved when the true number of clusters has been found, in which case the best performance is achieved with ARI $= 0.7398$ when $K = 75$ and $q = 0.9$. In the best case, the decision graph and the credal partition are illustrated in Figs. 7 and 8, respectively.

For comparison with this best case, we present in Fig. 9 the cluster centers selected by the degree of belief and the four typical densities mentioned in Remark 1. We can see that some

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2018.2869125, IEEE Transactions on Fuzzy Systems

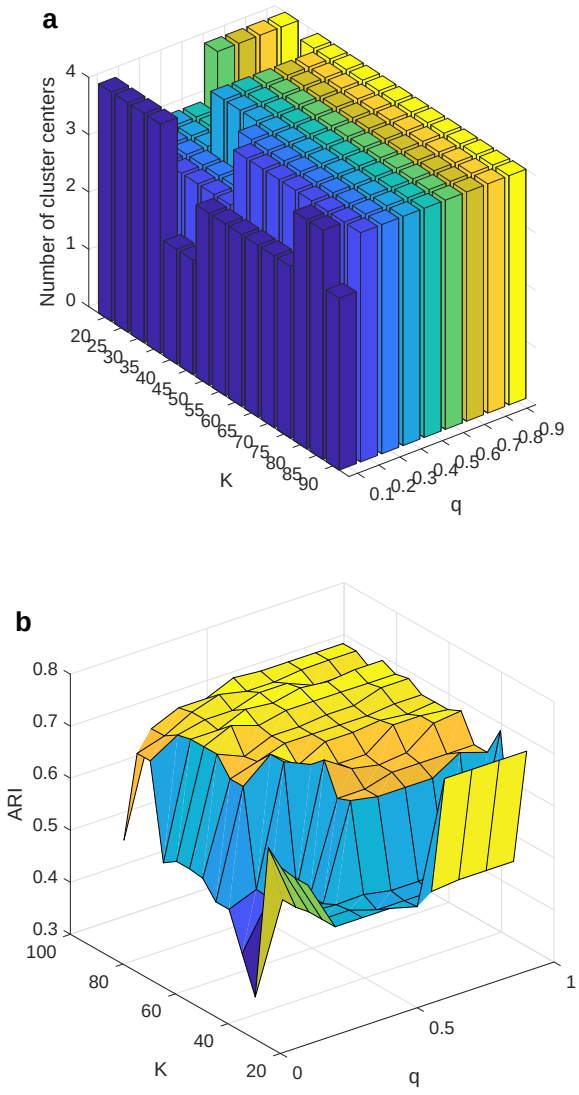

Fig. 6. Number of selected clusters (A) and ARI (b) as functions of K and q for the four-class dataset.

cluster centers selected by belief peaks are different from those obtained by density peaks. To study the influence of cluster centers, we computed the ARI for each group of selected cluster centers, by applying the same assignment strategy as used in DPC. With $K = 75$, $q = 0.9$, $f = 2\%$, we get $\mathrm{ARI}_{\mathrm{DPC}} = 0.3258$, $\mathrm{ARI}_{\mathrm{ADPC-KNN}} = \mathrm{ARI}_{\mathrm{FKNN-DPC}} = \mathrm{ARI}_{\mathrm{DPC-KNN}} = 0.7279$ and $\mathrm{ARI}_{\mathrm{BPC}} = 0.7392$, which, together with $\mathrm{ARI}_{\mathrm{BPEC}} = 0.7398$, shows that the cluster centers selected by belief peaks are better in this case. □

Example 5: In this example, we used the dataset *S2* in [31], consisting of 5000 objects and 15 clusters, to illustrate the partition via informative BPEC with a limited number of composite clusters. With $K = 80$, $q = 0.9$, $\Delta = 1$ and $\mathcal{K} = 1$, the decision graph is drawn in Fig. 10, and the rough approximations of the initial and final credal partitions are illustrated in Figs. 11 and 12, respectively. It can be seen from Fig. 10 that the true number of clusters can be found, and from Figs. 11 and 12 that the rough partition transformed from the

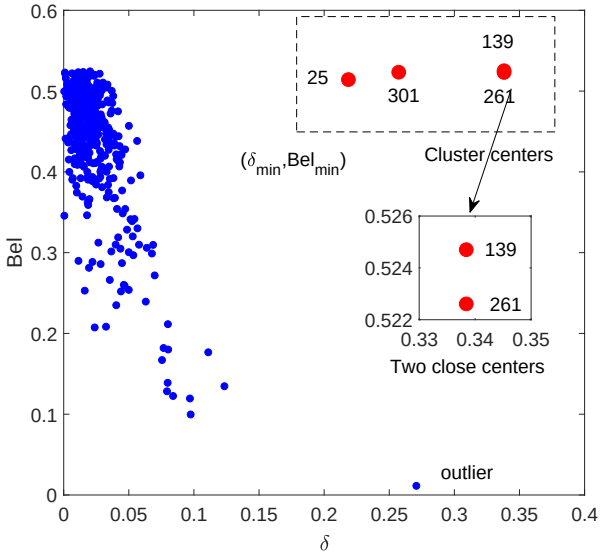

Fig. 7. $\delta - Bel$ decision graph for the four-class dataset.

credal partition obtained via informative BPEC seems to be reasonable. □

Remark 6: As can be seen from the decision graphs in above three examples that, the beliefs $Bel_i^{\mathcal{C}}(\{C\})$ seem to approach an upper bound that is smaller than 1. This is the case because, according to (15), we have

$$Bel_i^{\mathcal{C}}(\{C\}) \leq 1 - \lim_{d_{ij}\to 0} \prod_{j\in\mathcal{N}_K(o_i)} \left[1 - \phi(d_{ij}^2)\right] = 1 - (1-\alpha_0)^K. \quad (30)$$

When choosing $\alpha_0 = \frac{1}{K}$ we have $\lim_{K\to\infty}(1-\frac{1}{K})^K = \frac{1}{e}$ and thus get the infimum of upper bound, i.e., $1 - \frac{1}{e}$. To increase this infimum, α_0 can be redefined in a more general form such as $\alpha_0 = 1/K^\theta$ with a positive constant θ. With an appropriate choice of θ, correct number of clusters can be selected. □

B. Performance evaluation

In this subsection, we aim to evaluate the performances of BPEC based on some synthetic and UCI real-world datasets with characteristics summarized in Table II. As can be seen from Table II, the synthetic datasets have large numbers of clusters, while the real-word datasets have more attributes.

To implement BPEC, some parameters were preset as mentioned in Section III-D and some others were fixed individually as shown in Table III. In the absence of outliers, we suggest $\Delta = 1$ for simplicity. For comparison, several DPC algorithms were also applied to these datasets, including BPC, DPC [5], DPC-KNN [7], ADPC-KNN [9] and FKNN-DPC [10]. For DPC [5], the cutoff distance d_c was defined according to a proportion $f = 2\%$ of the total number of objects in a dataset, i.e., $d_c = \bar{d}(i_p)$, where $\bar{d}$ is a vector sorting distances $d_{ij}, i < j$ in descending order and $i_p = \mathsf{round}(n \cdot f)$. For DPC-KNN [7], $f = 2\%$. The ARI values are shown in Table IV.

As can be seen from Tables III and IV, BPEC and BPC can find the true number of clusters for most of these datasets,

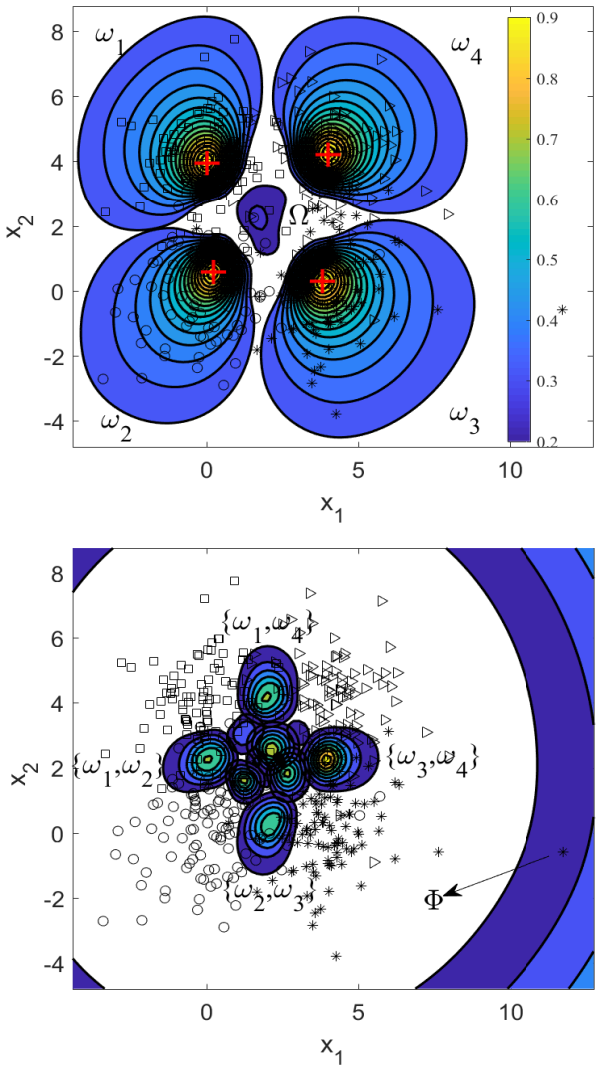

Fig. 8. Contour surfaces of the credal partition (top: singletons and Ω, bottom: empty set and composite clusters with $|A_j| = 2, 3$) obtained by BPEC for the four-class dataset.

TABLE II
DATASET DESCRIPTION

Type	Name	Size	# Attributes	c	Source
Synthetic	A3	7500	2	50	[32]
	D31	3200	2	31	[33]
	DIM1024	1024	1024	16	[34]
	R15	600	2	15	[33]
	S4	5000	2	15	[31]
	S2	5000	2	15	[31]
	Unbalance	650	2	8	[35]
Real-world	Iris	150	4	3	[36]
	Parkinsons	195	23	2	[36]
	Pima	768	8	2	[36]
	Seeds	210	7	3	[36]
	Waveform	5000	21	3	[36]
	WDBC	569	30	2	[36]
	Wine	178	13	3	[36]

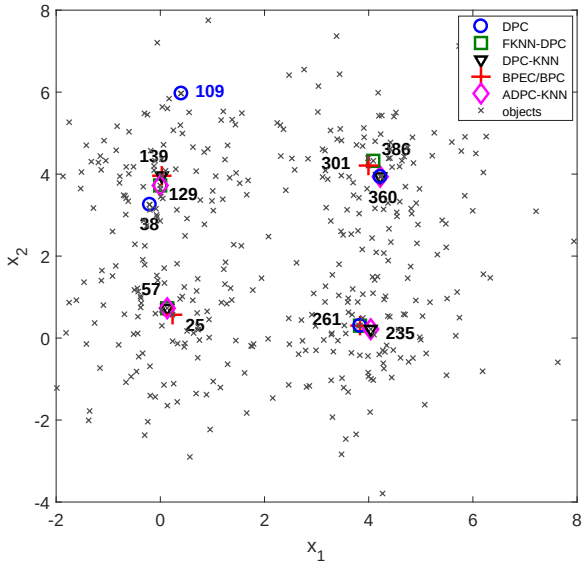

Fig. 9. Locations of centers selected by different methods for the four-class dataset.

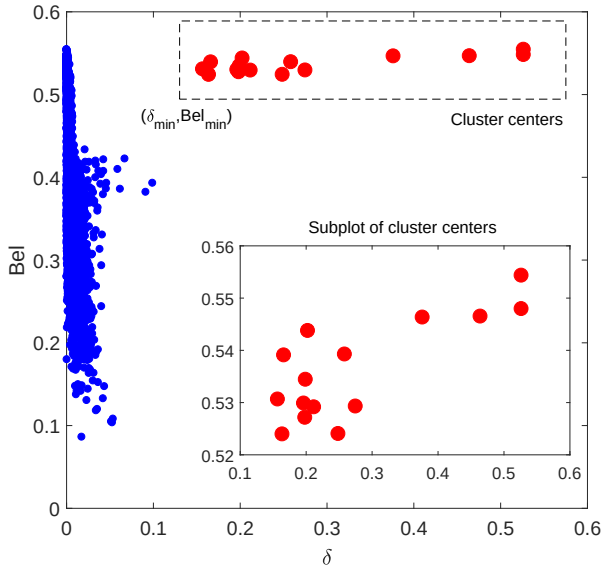

Fig. 10. $\delta - Bel$ decision graph for the S2 dataset

and BPEC has the highest ARI values in most cases. With the same fixed K and the same assignment strategy for DPC-KNN and BPC, we can see that BPC outperforms DPC-KNN and DPC in most cases. This result shows that cluster centers selected according to belief peaks can usually yield more better clustering performances. As stated in Remark 5, alternately tuning K and q in their ranges can usually induce more distinguishable decision graph and thus better performance. For instance, more appropriate cluster centers with ARI$= 0.8176$ can be achieved for the WDBC dataset when selecting $q = 0.1$ and $K = 40$ in an alternate way. As another example, the true number of clusters can be found for the Pima dataset with ARI$= 0.1210$ if $q = 0.1$ and $K = 40$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2018.2869125, IEEE Transactions on Fuzzy Systems

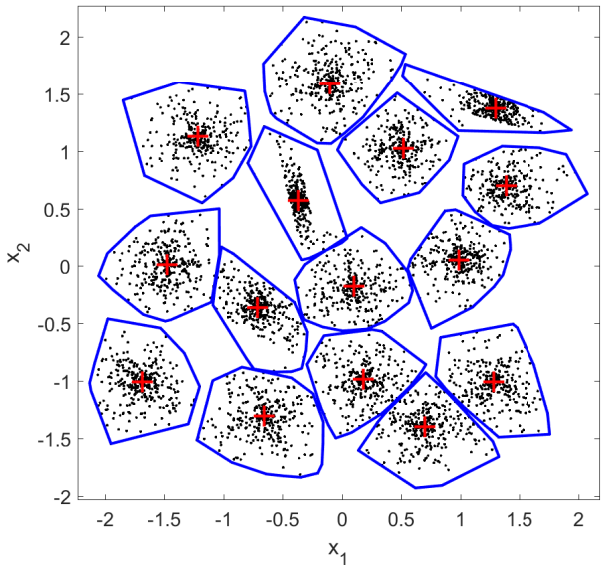

Fig. 11. Lower rough approximations of the initial credal partition $\mathcal{M}_0^{\Omega}$ for the S2 dataset.

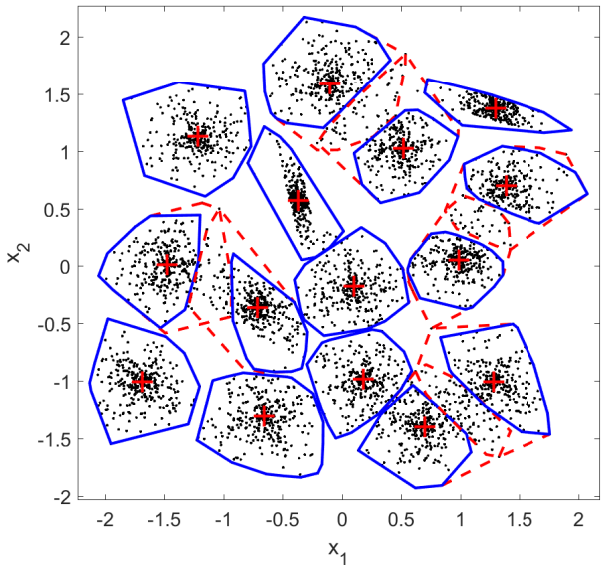

Fig. 12. Lower and (four pairs of) upper rough approximations of the final credal partition $\mathcal{M}^{\Omega}$ for the S2 dataset.

TABLE III
SELECTION OF SOME PARAMETERS IN BPEC

Datasets	Some parameters		Found clusters
	K	$\mathcal{K}$	by BPEC/BPC
A3	150	1	50 (50)
D31	50	1	31 (31)
DIM1024	20	1	16 (16)
R15	20	1	15 (15)
S4	300	1	15 (15)
S2	80	1	15 (15)
Unbalance	70	1	8 (8)
Iris	20	1	3 (3)
Parkinsons	40	1	2 (2)
Pima	40	1	2 (3)
Seeds	40	2	3 (3)
Waveform	100	2	3 (3)
WDBC	40	1	2 (2)
Wine	30	1	3 (3)

TABLE IV
ARI VALUES: COMPARISONS BETWEEN BPEC AND SOME ALTERNATIVE CLUSTERING ALGORITHMS [1, 2].

Datasets	ADPC -KNN	FKNN -DPC	DPC -KNN	DPC	BPC	BPEC
A3	0.97	–	0.9775	0.9246	0.9835	**0.9889**
D31	0.94	–	0.9358	0.8627	0.9384	**0.9522**
DIM1024	**1.00**	–	**1.00**	**1.00**	**1.00**	**1.00**
R15	0.99	–	**0.9928**	0.9228	**0.9928**	**0.9928**
S4	0.63	–	0.6268	0.5876	**0.6519**	0.6374
S2	–	0.924	0.9286	0.9227	0.8644	**0.9303**
Unbalance	**1.00**	–	**1.00**	**1.00**	**1.00**	**1.00**
Iris	0.76	**0.922**	0.7060	0.5681	0.7060	0.7565
Parkinsons	–	0.391	0.0266	0.3910	0.2566	**0.4135**
Pima	0.02	0.013	0	0.0143	0.0682	**0.0967**
Seeds	0.77	**0.790**	0.7076	0.6531	0.7076	0.7236
Waveform	0.25	0.350	0.2516	0.2669	0.2872	**0.3939**
WDBC	–	0.786	0.5175	0.5061	0.7546	**0.7924**
Wine	–	0.852	0.7128	0.6990	0.7269	**0.8653**

[1] The ARI values for ADPC-KNN and FKNN-DPC are taken from [9] and [10]. Missing values are indicated by "–".
[2] The bold and underlined value(s) in each row indicates the best performance.

By comparing BPEC with BPC, it can be concluded that ECM assigns the remaining objects more reliably than BPC, resulting in better performance. Furthermore, we can see that the hard partition obtained from the BPEC algorithm using the maximum plausibility-probability rule (9) is different from that obtained by the BPC algorithm, which means that BPC is not merely a crisp version of BPEC.

To compare BPEC with some other evidential clustering algorithms, ECM [15], EVCLUS [13] and EK-NNclus [37] were also implemented. When compared with ECM and EVCLUS, the BPEC algorithm was implemented with full focal sets. When compared with EK-NNclus, BPEC was used with singletons and the whole frame of discernment, as EK-NNclus. For ECM and EVCLUS, the number of clusters was preset to the value found by BPEC, and the Euclidean distance was used in BPEC instead of an adaptive one. In contrast with ECM and EVCLUS, EK-NNclus was run with integer labels randomly generated in the range $[1, n]$, as it does not require number of clusters as priori. Furthermore, the number of nearest neighbors (K_{nn}) and quantile of these nearest neighbors (q_{nn}) were set for EK-NNclus as follows, for the seven datasets from Iris to Wine: $(50, 0.5)$, $(300, 0.9)$, $(50, 0.9)$, $(500, 0.9)$, $(50, 0.5)$, $(200, 0.5)$ and $(50, 0.9)$.

For each real-world dataset in Table II, we run ECM, EVCLUS and EK-NNclus ten times, respectively. At each time, we calculated CRI between BPEC and these three

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2018.2869125, IEEE Transactions on Fuzzy Systems

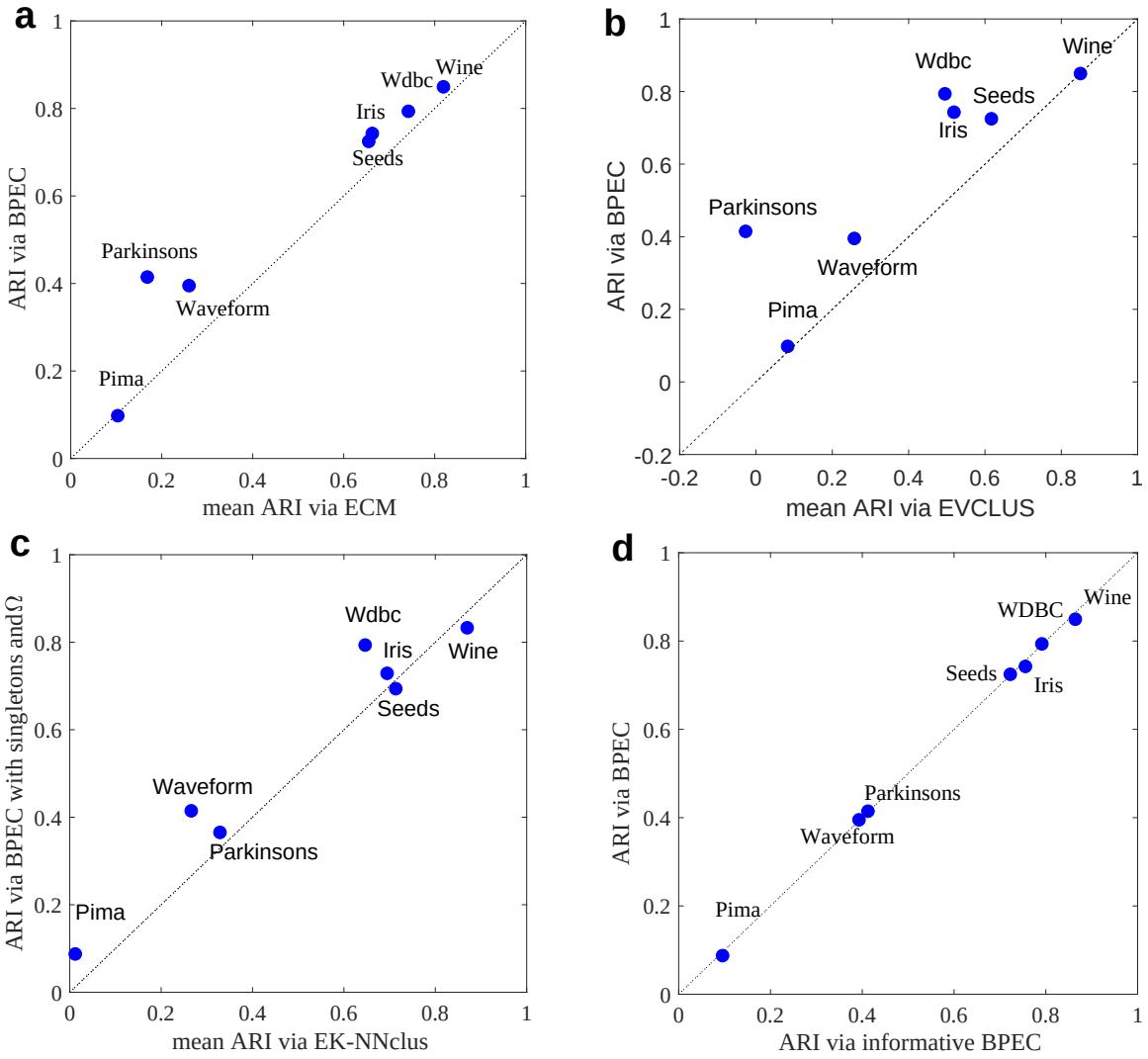

Fig. 13. The ARI values via BPEC vs. ARI values via (a) ECM, (b) EVCLUS, (c) EK-NNclus, and (d) informative BPEC with limited composite clusters

TABLE V
CRI ($mean \pm std.$) BETWEEN BPEC AND OTHER EVIDENTIAL CLUSTERING ALGORITHMS.

Datasets	BPEC-ECM	BPEC-EVCLUS	BPEC-EK-NNclus
Iris	0.9436 ± 0.0001	0.7407 ± 0.0222	0.3572 ± 0
Parkinsons	0.8434 ± 0.0001	0.5365 ± 0.0275	-0.1258 ± 0.0203
Pima	0.9331 ± 0.0005	0.6207 ± 0.0022	0.0045 ± 0
Seeds	0.9504 ± 0.0002	0.6053 ± 0.0047	0.3834 ± 0.0142
Waveform	0.9472 ± 0.0001	0.7157 ± 0.0016	0.3963 ± 0
WDBC	0.9596 ± 0.0003	0.6527 ± 0.0123	0.5352 ± 0
Wine	0.9616 ± 0.0001	0.7642 ± 0.0074	0.3423 ± 0

evidential algorithms. (Note that, for the EK-NNclus we only considered the cases when true number of clusters had been found). The mean CRI and standard deviation ($std.$) over the ten times are presented in Table V. It can be concluded from Table V that, on the one hand, the cluster centers found by the BPEC are not the same to those obtained by the ECM, on the other hand, BPEC creates different credal partitions from those generalized by the ECM, EVCLUS and EK-NNclus.

To gain further insight into the relative performances of BPEC and alternative evidential clustering algorithms, we compared the closeness of the hard partitions generated from credal partitions according to maximum rule (9) by each algorithm, to the true hard partition for each dataset. Fig. 13 displays the ARI values obtained by BPEC vs. those obtained by ECM, EVCLUS, and EK-NNclus. Fig. 13a shows that BPEC outperforms ECM on these seven real-world datasets. It can be seen from Fig. 13b that BPEC outperforms EVCLUS in most cases except for the Wine dataset, for which BPEC and EVCLUS have comparable performances. We can see from Fig. 13c that EK-NNclus outperforms BPEC on the Wine and Seeds datasets, whereas BPEC outperforms EK-NNclus on the other five real-world datasets. Finally, Fig. 13d shows that the performance of the BPEC algorithm is not deteriorated when limiting the number of composite clusters.

V. CONCLUSIONS

In this paper, we have introduced a clustering procedure, called the Belief-peaks Evidential Clustering (BPEC) algo-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2018.2869125, IEEE Transactions on Fuzzy Systems

rithm, which can find the true number of clusters and create a credal partition for some datasets with good performances. When the number of clusters is small (usually less than ten), the performances of BPEC and its informative variant with a limited number of composite clusters are approximately equal. In contrast, BPEC can be enhanced if less informative composite clusters (i.e., focal sets) when the number of clusters is large. Furthermore, BPEC can provide hard, fuzzy, possibilistic and even rough partitions. Finally, as a by-product of the BPEC algorithm, we proposed a belief version of DPC, called the Belief Peaks Clustering (BPC) algorithm. We have shown that BPC outperforms DPC in most cases but is outperformed by BPEC (of which it is not merely a crisp version).

There are several avenues for further research, such as combining the belief peak approach with other clustering algorithms instead of ECM, and improving the method to make it applicable to very large datasets.

ACKNOWLEDGEMENTS

The authors are grateful to the editor and the referees for the useful comments.

APPENDIX A
CALCULATIONS OF (25) AND (26) USING THE ALTERNATE OPTIMIZATION ALGORITHM IN CONSTRAINED ECM

The Lagrangian for problem (23)-(24) can be written as

$$\begin{aligned}\mathcal{L}(\mathcal{M}^{\Omega}, S_1, \ldots, S_c, \lambda_1, \cdots, \lambda_n, \eta_1, \ldots, \eta_c) \\ = \mathcal{J}_{BPEC} - \sum_{i=1}^{n} \lambda_i \Big(\sum_{j:A_j \neq \emptyset} m_{ij}^{\Omega} + m_{i\emptyset}^{\Omega} - 1 \Big) \\ - \sum_{k=1}^{c} \eta_k \Big(\det(S_k) - 1 \Big), \quad (31)\end{aligned}$$

where λ_i and η_k are Lagrange multipliers.

First, we optimize m_{ij}^{Ω} and $m_{i\emptyset}^{\Omega}$ by fixing matrix S_k. By calculating the partial derivatives of (31) with respect to m_{ij}^{Ω}, $m_{i\emptyset}^{\Omega}$ and λ_i and setting them to zero, we have for $i = 1, 2, \ldots, n$ and all j such that $A_j \subseteq \Omega$,

$$\frac{\partial \mathcal{L}}{\partial m_{ij}^{\Omega}} = \beta |A_j|^{\alpha} (m_{ij}^{\Omega})^{\beta - 1} D_{ij}^2 - \lambda_i = 0, \quad (32a)$$

$$\frac{\partial \mathcal{L}}{\partial m_{i\emptyset}^{\Omega}} = \beta \Delta^2 (m_{i\emptyset}^{\Omega})^{\beta - 1} - \lambda_i = 0, \quad (32b)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = \sum\nolimits_{j:A_j \neq \emptyset} m_{ij}^{\Omega} + m_{i\emptyset}^{\Omega} - 1 = 0. \quad (32c)$$

By solving m_{ij}^{Ω} and $m_{i\emptyset}^{\Omega}$ from (32a) and (32b) and inserting them into (32c), we get

$$\Big(\frac{\lambda_i}{\beta} \Big)^{\frac{1}{(\beta-1)}} = \frac{1}{\sum\limits_{j:A_j \neq \emptyset} |A_j|^{-\alpha/(\beta-1)} D_{ij}^{-2/(\beta-1)} + \Delta^{-2/(\beta-1)}}, \quad (33)$$

which, inserted in (32a) and (32b), leads to (25).

Next, we consider that m_{ij}^{Ω} and $m_{i\emptyset}^{\Omega}$ are fixed to obtain S_k and η_k. We have, for $k = 1, 2, \ldots, c$,

$$\frac{\partial \mathcal{L}}{\partial S_k} = \sum_{i=1}^{n} \sum_{j:A_j \neq \emptyset} |A_j|^{\alpha - 1} (m_{ij}^{\Omega})^{\beta} s_{kj} (x_i - \bar{v}_j)(x_i - \bar{v}_j)' - \eta_k \det(S_k) S_k^{-1} = 0, \quad (34)$$

and

$$\frac{\partial \mathcal{L}}{\partial \eta_k} = \det(S_k) - 1 = 0. \quad (35)$$

Define

$$\Sigma_k = \sum_{i=1}^{n} \sum_{j:A_j \neq \emptyset} |A_j|^{\alpha - 1} (m_{ij}^{\Omega})^{\beta} s_{kj} (x_i - \bar{v}_j)(x_i - \bar{v}_j)'.$$

From (34) and (35), it follows that

$$\Sigma_k S_k = \eta_k I, \quad k = 1, 2, \cdots, c, \quad (36)$$

where I is a $p \times p$ identify matrix. Taking the determinant of (36), we get

$$\det(\Sigma_k S_k) = \det(\Sigma_k) \det(S_k) = \det(\Sigma_k) = \eta_k^p, \quad (37)$$

which leads to

$$\eta_k = \det(\Sigma_k)^{1/p}. \quad (38)$$

By inserting (38) into (34), we finally get (26).

APPENDIX B
BPC: BELIEF PEAKS CLUSTERING ALGORITHM

The belief version of DPC algorithm, i.e., belief peaks clustering algorithm (BPC) is summarized as follow.

Algorithm 2: Belief peaks clustering

Input: K, α_0, q, $x_i \in \mathbb{R}^p$ for $i = 1, 2, \cdots, n$

1 Calculate degrees of belief ($Bel_i^{\mathcal{C}}(\{\mathrm{C}\})$) for all objects using (15)
2 Calculate delta's (δ_i) for all objects according to (18)
3 Draw the decision graph $\delta - Bel$
4 Select cluster centers $v_k, k = 1, 2, \cdots c$
5 Assign each of the remaining objects to the same cluster as its nearest neighbor with higher belief

Output: cluster centers and hard partition

REFERENCES

[1] C. Bouveyron and C. Brunet-Saumard, "Model-based clustering of high-dimensional data: A review," *Computational Statistics and Data Analysis*, vol. 71, no. 1, pp. 52–78, 2014.

[2] T. Denoeux and O. Kanjanatarakul, "Evidential clustering: a review," in *5th International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM)*, Da Nang, Dec. 2016, pp. 24–35.

[3] A. Gosain and S. Dahiya, "Performance analysis of various fuzzy clustering algorithms: A review," in *Proceedings of International conference on communication, computing and virtualization*, Mumbai, Indian, Feb. 2016, pp. 100–111.

[4] A. Saxena, M. Prasad, A. Gupta, and et. al., "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[5] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, pp. 1492–1496, 2014.

[6] V. Courjault-Rade, L. D'Estampes, and S. Puechmorel, "Improved density peak clustering for large datasets," 2016, hAL Id: hal-01353574.
[7] M. Du, S. Ding, and H. Jia, "Study on density peaks clustering based on k-nearest neighbors and principal component analysis," *Knowledge-Based Systems*, vol. 99, pp. 135–145, 2016.
[8] Y. Geng, Q. Li, R. Zheng, F. Zhuang, R. He, and N. Xiong, "RECOME: A new density-based clustering algorithm using relative KNN kernel density," *Information Sciences*, vol. 436-437, pp. 13–30, 2018.
[9] Y. Liu, Z. Ma, and F. Yu, "Adaptive density peak clustering based on k-nearest neighbors with aggregating strategy," *Knowledge-Based Systems*, vol. 133, pp. 208–220, 2017.
[10] J. Xie, H. Gao, W. Xie, X. liu, and P. W. Grant, "Robust clustering by detecting density peaks and assigning points based on fuzzy weighted k-nearest neighbors," *Information Sciences*, vol. 354, pp. 19–40, 2016.
[11] V. Antonie, B. Quost, M.-H. Masson, and T. Denoeux, "CECM: Constrained evidential c-means algorithm," *Computational Statistics and Data Analysis*, vol. 56, no. 4, pp. 894–914, 2012.
[12] Z.-G. Liu, J. Dezert, G. Mercier, and Q. Pan, "Belief C-Means: An extension of fuzzy C-Means algorithm in belief functions framework," *Pattern Recognition Letter*, vol. 33, pp. 291–300, 2012.
[13] T. Denoeux and M.-H. Masson, "EVCLUS: evidential clustering of proximity data," *IEEE Transactions on Systems, Man and Cybernetics B*, vol. 34, no. 1, pp. 95–109, 2004.
[14] T. Denoeux, S. Sriboonchitta, and O. Kanjanatarakul, "Evidential clustering of large dissimilarity data," *Knowledge-Based Systems*, vol. 106, pp. 195–219, 2016.
[15] M.-H. Masson and T. Denoeux, "ECM: an evidential version of the fuzzy c-means algorithm," *Pattern Recognition*, vol. 41, no. 4, pp. 1364–1397, 2008.
[16] K. Zhou, A. Martin, Q. Pan, and Z.-G. Liu, "Median evidential c-means algorithm and its application to community detection," *Knowledge-Based Systems*, vol. 74, no. 1, pp. 69–88, 2015.
[17] G. Shafer, *A mathematical theory of evidence*. Princeton, N.J.: Princeton University Press, 1976.
[18] A. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Annals of Mathematical Statistics*, vol. 38, pp. 325–339, 1987.
[19] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, pp. 191–243, 1994.
[20] J. Bezdek, "Pattern recognition with fuzzy objective function algorithm," 1981.
[21] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, vol. 1, pp. 98–111, 1993.
[22] P. Lingras and G. Peters, "Applying rough set concepts to clustering," in *Rough Sets: Selected Methods and Applications in Management and Engineering*, G. Peters, P. Lingras, D. Slezak, and Y. Yao, Eds. London, UK: Springer-Verlag, 2012, pp. 23–37.
[23] G. Peters, "Is there any need for rough clustering?" *Pattern Recognition Letter*, vol. 53, pp. 31–37, 2015.
[24] T. Denoeux, "Conjunctive and disjunctive combination of belief functions induced by non-distinct bodies of evidence," *Artificial Intelligence*, vol. 172, no. 2-3, pp. 234–264, 2008.
[25] ——, "40 years of Dempster-Shafer theory," *International Journal of Approximate Reasoning*, vol. 79, pp. 1–6, 2016.
[26] G. Shafer, "A mathematical theory of evidence turns 40," *International Journal of Approximate Reasoning*, vol. 79, pp. 7–25, 2016.
[27] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 5, pp. 804–813, 1995.
[28] Z.-G. Su, T. Denoeux, Y.-S. Hao, and M. Zhao, "Evidential K-NN classification with enhanced performance via optimizing a class of parametric t-rules," *Knowledge-Based Systems*, vol. 142, pp. 7–16, 2018.
[29] W. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.
[30] T. Denoeux, S. Li, and S. Sriboonchitta, "Evaluating and comparing soft partitions: an approach based on Dempster-Shafer theory," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 3, pp. 1231–1244, 2018.
[31] P. Franti and O. Virmajoki, "Iterative shrinking method for clustering problems," *Pattern Recognition*, vol. 39, no. 5, pp. 761–765, 2006.
[32] I. Karkkainen and P. Franti, "Dynamic local search algorithm for the clustering problem," Tech. Rep., 2002, a-2002-6.
[33] C. Veenman, M. Reinders, and E. Backer, "A maximum variance cluster algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1273–1280, 2002.
[34] P. Franti, O. Virmajoki, and V. Hautamaki, "Fast agglomerative clustering using a k-nearest neighbor graph," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1875–1881, 2006.
[35] M. Rezaei and P. Franti, "Set-matching methods for external cluster validity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2173–2186, 2016.
[36] K. Bache and M. Lichman, "UCI machine learning repository," 2013, uRL http://archive.ics.uci.edu/ml.
[37] T. Denoeux, O. Kanjanatarakul, and S. Sriboonchitta, "EK-NNclus: A clustering procedure based on the evidential k-nearest neighbor rule," *Knowledge-Based Systems*, vol. 88, pp. 57–69, 2015.

Zhi-gang Su Zhi-gang Su received his M.S. and Ph.D from Southeast University (SEU), China, in 2006 and 2010 respectively, and then became an assistant professor with the Dept. of Energy Information and Automation, School of Energy and Environment at the SEU. In 2013, he became an associate professor. From 2014 to 2015, he worked as a visiting scholar with Dept. of Electrical and Computer Engineering at The University of Texas at San Antonio in USA. His research interests concern artificial intelligence and theory of belief function with applications to pattern recognition, data mining and in particular to similar practical issues in thermal power engineering. He is also interested in nonlinear control theory with applications to thermal processes, and he was selected as one outstanding reviewer by the journal *Automatica* in 2016-2017. He is the first author of more than 20 papers in journals including *IEEE Transactions on Fuzzy Systems*, *IEEE Transactions on Automatic Control*, *Automatica*, *International Journal of Approximate Reasoning* and *Engineering Applications of Artificial Intelligence*.

Thierry Denoeux Thierry Denoeux is a Full Professor (Exceptional Class) with the Department of Information Processing Engineering at the Compiègne University of Technology (UTC), France. He has been deputy director of the Heudiasyc research Lab (UMR 7253) from 2008 to March 2014 and a Vice President of the Scientific Council of UTC in 2012-2014. He is the scientific coordinator of the Laboratory of Excellence Technological Systems of Systems. From 2016 to 2018, he has been an Overseas Talent visiting professor at Beijing University of Technology. His research interests concern reasoning and decision-making under uncertainty and, more generally, the management of uncertainty in intelligent systems. His main contributions are in the theory of belief functions with applications to statistical inference, pattern recognition, machine learning and information fusion. He is the author of more than 200 papers in journals and conference proceedings and he has supervised 30 PhD theses. He is the Editor-in-Chief of the *International Journal of Approximate Reasoning* (Elsevier), and an Associate Editor of several journals including *Fuzzy Sets and Systems* and *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*.