Evidential clustering for categorical data

1st Abdoul Jalil Djiberou Mahamadou
LIMOS, UMR 6158, F-63000
Clermont Auvergne University
Clermont-Ferrand, France
abdoul_jalil.djiberou_mahamadou@uca.fr

2nd Violaine Antoine
LIMOS, UMR 6158, F-63000
Clermont Auvergne University
Clermont-Ferrand, France
violaine.antoine@uca.fr

3rd Gregory J. Christie
Digital Health Hub
Simon Fraser University
Vancouver, Canada
greg_christie@sfu.ca

4th Sylvain Moreno
Digital Health Hub
Simon Fraser University
Vancouver, Canada
sylvain_moreno@sfu.ca

***Abstract*—Evidential clustering methods assign objects to clusters with a degree of belief, allowing for better representation of cluster overlap and outliers. Based on the theoretical framework of belief functions, they generate credal partitions which extend crisp, fuzzy and possibilistic partitions. Despite their ability to provide rich information about the partition, no evidential clustering algorithm for categorical data has yet been proposed. This paper presents a categorical version of ECM, an evidential variant of *k-means*. The proposed algorithm, referred to as cat-ECM, considers a new dissimilarity measure and introduces an alternating minimization scheme in order to obtain a credal partition. Experimental results with real and synthetic data sets show the potential and the efficiency of cat-ECM for clustering categorical data.**

***Index Terms*—clustering, categorical data, credal partition, evidential c-means, belief functions**

I. Introduction

Clustering is a fundamental data mining technique that aims, without any other prior information, to group objects on the basis of a similarity notion. In such unsupervised contexts, objects are usually described by numerical attributes and the similarity notion corresponds to a distance measured between pairs of objects or between objects and clusters. Numerous clustering approaches have been proposed in the literature in order to meet the specific needs of various problems [1]. Among efficient clustering techniques, the family of partition-based methods, including the popular *k-means* algorithm, have been widely used because they are relatively simple to compute, easy to interpret, and they scale efficiently with large data sets. Accordingly, partition-based clustering methods have seen widespread use in disparate fields including bioinformatics with gene expression [2], robotics with data sensors analysis [3], business [4], climatology [5], and others.

The *k-means* algorithm is a partition-based clustering method which represents clusters by a centroid surrounded by a crisp partition, with objects belonging unambiguously to a single cluster and not to any other. In many real-world applications however, inter-object differences can be ambiguous or uncertain, and the use of crisp partitions can lead to poor overall classification accuracy under such conditions. To address this problem and to capture this degree of ambiguity, soft clustering variants have been proposed that allow the expression of uncertainty or/and imprecision in the partition. The *fuzzy c-means* (FCM) algorithm [6], based on the probabilistic theory, provides a fuzzy partition where each object has a degree of membership to each cluster. Since FCM has poor robustness against noise and outliers, possibilistic extensions of *k-means* [7], [8] as well as a variant of FCM called *noise-clutering* (NC) [9] has been introduced. More recently, an evidential clustering version of NC, referred to as *evidential c-means* (ECM), has been proposed [10]. By using the belief functions theory [11] and by generating a credal partition, ECM enhances hard, fuzzy and possibilistic partitions [12], while expressing with more precision the magnitude of doubt concerning the class membership of the various objects.

Initially, *k-means* and its variants were created for the clustering of numerical data. These types of data allow the use of geometric distances (e.g. Euclidean) to define similarities between objects. However, many real-world data sets include qualitative variables which do not have geometrical properties and which cannot be analyzed with clustering algorithms that rely on geometrical distances. To address this limitation, several new extensions of *k-means* have been proposed [13]–[17]. In [13], the standard *k-means* algorithm is adjusted to use a hard centroid representation and to handle a dissimilarity measure fitting categorical objects. The algorithm, which generates a crisp partition, is generalized by [14] to produce a fuzzy partition. Later, [15] proposed to improve the categorical fuzzy clustering by defining fuzzy centroids. More recent works include categorical *k-means* versions with effective dissimilarity functions [16], [17].

Although the generation of a credal partition via an evidential clustering allows the expression of a wide variety of situations ranging from complete ignorance to full certainty, there currently exists no evidential clustering algorithms dedicated to categorical data. In the present work, a new variant of *k-means* that takes as inputs qualitative variables and which generates a credal partition is introduced. The remainder of this paper is organized as follows: Section II recalls the necessary background in clustering and in belief function theory. Section III details the new categorical evidential clustering algorithm, then Section IV illustrates the applicability of the method on two well-known clustering data sets. The final section presents the conclusions and suggests future extensions for the work.

II. Clustering preliminaries

A. Fuzzy c-means variants

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a collection of n objects where $\mathbf{x}_i = [x_{i1}, \ldots, x_{il}, \ldots, x_{ip}]$ is a vector of p observed features

for the i^{th} object. Thus, x_{il} denotes the value of the l^{th} feature for the object $\mathbf{x}_i$. The *fuzzy c-means* (FCM) variants aim at grouping objects into c clusters characterized by prototypes (or centroids). Let $\mathbf{v}_k = [v_{k1}, \ldots, v_{kl}, \ldots, v_{kp}]$ be the p-dimensional prototype of the k cluster and $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_c\}$ be the set of c prototypes. All FCM variants generate a fuzzy partition $\mathbf{U} = [u_{ik}]$, where u_{ik} corresponds to the degree of membership between the object i and the cluster k, by minimizing the following objective function:

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

such that $\sum_{k=1}^{c} u_{ik} = 1$ and $u_{ik} \geq 0$ for all $i = \{1, \ldots n\}$ and $k = \{1, \ldots c\}$.

The β coefficient is a fixed parameter that controls the fuzziness of the partition and d_{ik} is a dissimilarity measure computed between object $\mathbf{x}_i$ and centroid $\mathbf{v}_k$. Historically, FCM is dedicated to group objects whose features are given by numerical values [6]. Since the minimization of the objective function is a NP-hard problem, a heuristic given a local minimum is employed. It consists of performing an iterative optimization of the fuzzy partition $\mathbf{U}$ and the centroids $\mathbf{V}$.

Later, the FCM algorithm was adapted in order to take into account qualitative data [14], [15]. Let $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_l, \ldots, \mathbf{A}_p)$ be a set of p categorical attributes and $Dom(\mathbf{A}_l) = (a_l^{(1)} \ldots a_l^{(n_l)})$ be the domain of possible values for the attribute $\mathbf{A}_l$. Such attributes are not ordered and contain a finite number of values, e.g n_l for $\mathbf{A}_l$. In [14], centroids are hard, i.e. they are defined as objects with categorical values. The algorithm, referred to as *fuzzy k-modes* (FKM), introduces the following dissimilarity measure:

$$d_{ik}^2 = \sum_{l=1}^{p} \phi(x_{il}, v_{kl}), \tag{2}$$

such that $\phi(x_{il}, v_{kl}) = 1$ if $x_{il} \neq v_{kl}$ and equals 0 otherwise.

Although FKM has relatively good performances, the mixing of a fuzzy partition with hard centroids is arguably questionable: FKM permits doubt on cluster assignments, but forces prototypes to have a single attribute value for each attribute. Consequently, other attribute values with a high frequency (but not the largest one) are ignored outright. To resolve this issue, [15] introduced the notion of categorical fuzzy centroids. For each feature l of each cluster k, they defined weights w_{kl} such that $v_{kl} = [w_{kl}^{(1)} a_l^{(1)} \wedge \cdots \wedge w_{kl}^{(n_l)} a_l^{(n_l)}]$ and $w_{kl}^{(t)} \geq 0$ is the weight of the t^{th} value in the domain $\mathbf{A}_l$ for the cluster k. The weights w_{kl} should respect the following constraint:

$$\sum_{t=1}^{n_l} w_{kl}^{(t)} = 1, \quad \forall k \in \{1 \ldots c\}, l \in \{1 \ldots p\}. \tag{3}$$

Then, a distance between a fuzzy centroid $\mathbf{v}_k$ and an object $\mathbf{x}_i$ is calculated by summing the weight of attribute values that are different from the attribute values of $\mathbf{x}_i$:

$$d_{ik}^2 = \sum_{l=1}^{p} \phi'(x_{il}, v_{kl}), \tag{4}$$

such that

$$\phi'(x_{il}, v_{kl}) = \sum_{t \in \mathcal{D}_{il}} w_{kl}^{(t)}, \tag{5}$$

with $\mathcal{D}_{il} = Dom(\mathbf{A}_l) \setminus a_l^{(r)}$ and $a_l^{(r)} = x_{il}$.

B. Belief functions

Belief functions theory corresponds to the Dempster-Shafer theory of evidence [11], [18], which defines a mathematical framework for modeling partial and unreliable information. Let us consider a variable ω taking values in a finite set $\Omega = \{\omega_1, \ldots, \omega_c\}$ called the frame of discernment. The mass function $m : 2^{\Omega} \rightarrow [0, 1]$ represents the partial knowledge regarding the actual value taken by ω. It satisfies $\sum_{A \subseteq \Omega} m(A) = 1$.

Any subset A such that $m(A) > 0$ is called a focal set of m. Complete ignorance is represented by $m(\Omega) = 1$ and full certainty is obtained when a unique singleton of Ω possesses the whole mass of belief.

Several operations and measures have been proposed in order to simplify the interpretation of a mass function and to make a decision regarding the value of ω. For example, the *pignistic transformation* allows one to convert a mass function to a probability distribution [18] and the normalized *non-specificity* measure $N^{\star}$ evaluates the degree of imprecision of a mass function m, allowing one to quantify the degree of information included in m [10], [19].

C. Evidential c-means

The evidential c-means (ECM) [10] is a generalization of the FCM algorithm dedicated for numerical data. It generates a credal partition $\mathbf{M} = (m_i)$ in which a mass function m_i is defined for each object i. The use of evidential theory allows the representation of uncertainties and imprecision regarding the class membership of the objects. Since any subsets A_j in the set $\Omega = \{\omega_1, \ldots, \omega_c\}$ of possible classes can be a focal set, the ECM algorithm represents not only prototypes for clusters, but also prototypes for subsets with a cardinality higher than 1. Then, for each subset $A_j \subseteq \Omega$, $A_j \neq \emptyset$, a centroid $\mathbf{v}_j \in \mathbb{R}^p$ is calculated as the barycenter of the centers associated to each class of A_j.

The ECM algorithm searches for a credal partition $\mathbf{M}$ and the set of prototypes $\mathbf{V}$ that minimize intra-cluster variance:

$$J_{ECM}(\mathbf{M}, \mathbf{V}) = \sum_{i=1}^{n} \sum_{A_j \subseteq \Omega, A_j \neq \emptyset} |A_j|^{\alpha} m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta}, \tag{6}$$

such that, for all $i = \{1, \ldots n\}$ and for all $A_j \subseteq \Omega$,

$$\sum_{j/A_j \subseteq \Omega, A_j \neq \emptyset} m_{ij} + m_{i\emptyset} = 1, \qquad m_{ij} \geq 0. \tag{7}$$

The mass $m_{i\emptyset}$ denotes the mass of $\mathbf{x}_i$ allocated to the empty set, $\rho > 0$ is a fixed parameter allowing the user to control the importance given to the empty set, and d_{ij} represents the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{v}_j$. The weighting coefficient $|A_j|^{\alpha}$, which corresponds to the cardinality of A_j as a power of α, allows the user to penalize the allocation of

belief to subsets with high cardinality. As in FCM, $\beta > 1$ corresponds to an exponent that controls the fuzziness of the partition: β close to 1 gives a credal partition similar to a crisp partition, whereas β with a high value provides a partition where coefficients are equally distributed throughout the clusters. Usually, α is set to 1 and β to 2.

III. CATEGORICAL ECM

In this section, we introduce a new evidential algorithm referred to as cat-ECM. The method is a variant of ECM that takes in account qualitative attributes.

A. Notations and Objective function

Similarly to FCM with fuzzy centroids, let us assume that the collection of objects to be clustered is defined by a set of categorical features $\mathbf{A}$. For each subset $A_j \subseteq \Omega$, $A_j \neq \emptyset$, we introduce a fuzzy prototype $\mathbf{v}_j = (v_{j1}, \ldots, v_{jl}, \ldots, v_{jp})$ such that $v_{jl} = [w_{jl}^{(1)} a_l^{(1)} \wedge \cdots \wedge w_{jl}^{(n_l)} a_l^{(n_l)}]$.

The weight $w_{jl}^{(t)}$ is a positive value that corresponds to the coefficient given to the *tth* value in the domain $\mathbf{A}_l$ for the subset A_j. We propose to associate to this weight the barycenter of the weight of the classes composing A_j:

$$w_{jl}^{(t)} = \frac{1}{|A_j|} \sum_{\omega_k \in A_j} w_{kl}^{(t)}. \tag{8}$$

By constraining for each cluster the sum of the weight to 1 (cf. Eq. (3)), we obtain, for subsets A_j such that $|A_j| > 1$, the following rule:

$$\sum_{t=1}^{n_l} w_{jl}^{(t)} = 1 \quad \forall l \in \{1 \ldots p\}, \forall A_j \subseteq \Omega, A_j \neq \emptyset. \tag{9}$$

Then, we define the squared dissimilarity measure d_{ij}^2 between an object $\mathbf{x}_i$ and a subset A_j as

$$d_{ij}^2 = \frac{1}{p} \sum_{l=1}^{p} \phi''(x_{ij}, v_{jl}), \tag{10}$$

where

$$\phi''(x_{ij}, v_{jl}) = \sum_{t \in \mathcal{D}_{il}} w_{jl}^{(t)} = \sum_{t \in \mathcal{D}_{il}} \frac{1}{|A_j|} \sum_{\omega_k \in A_j} w_{kl}^{(t)}. \tag{11}$$

The function d_{ij}^2 provides a normalized dissimilarity value by considering only attribute values different from the object value.

The goal of cat-ECM is to create a credal partition $\mathbf{M}$ with the best set of weights $\mathbf{W}$ that minimize (6) such that (3) and (7) are respected. Note that weights for subsets A_j such that $|A_j| > 1$ are defined by weights associated to singletons (cf. Eq. (8)). Thus, the set $\mathbf{W}$ only includes singletons weights.

B. Optimization

Minimizing $J_{ECM}(\mathbf{M}, \mathbf{W})$ can be solved by iteratively optimizing $\mathbf{M}$ and $\mathbf{W}$ until convergence.

First, the weights (and consequently the centroids) are fixed and the objective function is minimized with the respect to the credal partition $\mathbf{M}$ and subject to conditions (7). In this framework, d_{ij}^2 is considered as a fixed coefficient. Hence, the constrained problem is identical to ECM and can be solved by introducing Lagrange multipliers [10]. The update formula of m_{ij} is, $\forall i = 1, \ldots n$, $\forall j / A_j \subseteq \Omega$ and $A_j \neq \emptyset$:

$$m_{ij} = \frac{|A_j|^{-\alpha/(\beta-1)} d_{ij}^{-2/(\beta-1)}}{\sum_{A_k \neq \emptyset} |A_k|^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)} + \delta^{-2/(\beta-1)}}. \tag{12}$$

For $A_j = \emptyset$, the mass is defined as:

$$m_{i\emptyset} = 1 - \sum_{A_k \neq \emptyset} m_{ij} \quad \forall i = 1, \ldots n. \tag{13}$$

Second, the credal partition is fixed and J_{ECM} is minimized with respect to the set of weights $\mathbf{W}$ and subject to constraint (3). Since each set of weights $w_{kl} = \{w_{kl}^{(1)}, \ldots, w_{kl}^{(n_l)}\}$ are independent $\forall k \in \{1 \ldots c\}$ and $\forall l \in \{1 \ldots p\}$, minimizing the objective function is equivalent to minimizing each w_{kl} separately. The problem is a linear optimization problem that is solved by giving the maximal weight to the categorical value that is the most frequent in the cluster. The update formula obtained for the weight associated to cluster k, attribute l and the *tth* possible value of the attribute is

$$w_{kl}^{(t)} = \begin{cases} 1 \text{ if} & f_{ik}^{(t)} > f_{ik}^{(r)}, \forall r \in \{1 \ldots n_l\}, r \neq t, \\ 0 \text{ if} & f_{ik}^{(t)} < f_{ik}^{(r)}, \text{ s.t. } r \in \{1 \ldots n_l\}, r \neq t, \\ \frac{1}{q} \text{ if} & f_{ik}^{(t)} = f_{ik}^{(s_1)} = \cdots = f_{ik}^{(s_{q-1})} > f_{ik}^{(r)}, \\ & \forall s_1, \ldots, s_{q-1}, r \in \{1 \ldots n_l\}, \\ & r \neq s_1 \neq \cdots \neq s_{q-1} \neq t, \end{cases} \tag{14}$$

where

$$f_{ik}^{(t)} = \frac{1}{p} \sum_{A_j \supseteq \omega_k} \sum_{\mathcal{S}_l^{(t)}} |A_j|^{\alpha-1} m_{ij}^{\beta}, \tag{15}$$

and $\mathcal{S}_l^{(t)}$ defines a subset of objects in $\{1 \ldots n\}$ such that $x_{il} = a_l^{(t)}$.

Proof. Let b_{ij} define a scalar such that $b_{ij} \triangleq \frac{1}{p} |A_j|^{\alpha-1} m_{ij}^{\beta}$, $\forall i = 1, n \quad \forall j / A_j \subseteq \Omega, A_j \neq \emptyset$. Replacing d_{ij}^2 in the objective function (6) by (10), (11) and b_{ij} gives

$$J_{ECM}(\mathbf{M}, \mathbf{W}) = \sum_{i=1}^{n} \sum_{\substack{A_j \subseteq \Omega, \\ A_j \neq \emptyset}} b_{ij} \sum_{l=1}^{p} \sum_{t \in D_{il}} \sum_{\omega_k \in A_j} w_{kl}^{(t)} + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta}.$$

Since $\sum_{A_j \subseteq \Omega} \sum_{\omega_k \in A_j} b_{ij} = \sum_{k=1}^{c} \sum_{A_j \subseteq \Omega, \omega_k \in A_j} b_{ij}$, the objective function can be written as:

$$J_{ECM}(\mathbf{M}, \mathbf{W}) = \sum_{k=1}^{c} \sum_{l=1}^{p} \sum_{i=1}^{n} \sum_{\substack{A_j \subseteq \Omega, \\ \omega_k \in A_j}} b_{ij} \sum_{t \in D_{il}} w_{kl}^{(t)} + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta}.$$

Let $w_{kl}^{(r)}$ be the weight associated to the attribute value egal to x_{il}. Using (3), we deduce that $\sum_{t \in D_{il}} w_{kl}^{(t)} = 1 - w_{kl}^{(r)}$. Thus,

$$J_{ECM}(\mathbf{M}, \mathbf{W}) = \sum_{k=1}^{c} \sum_{l=1}^{p} \sum_{i=1}^{n} \sum_{\substack{A_j \subseteq \Omega, \\ \omega_k \in A_j}} b_{ij} (1 - w_{kl}^{(r)}) + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta}.$$

By fixing $\mathbf{M}$ to minimize J_{ECM}, the terms $\sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta}$ and $\sum_{i=1}^{n} \sum_{A_j \subseteq \Omega, \omega_k \in A_j} b_{ij}$ become constants. Since each element w_{kl} are independent, minimizing $J_{ECM}(\mathbf{W})$ is equivalent to maximizing (16) under the same conditions.

$$J'_{ECM}(w_{kl}) = \sum_{i=1}^{n} \sum_{A_j \subseteq \Omega, \omega_k \in A_j} b_{ij} w_{kl}^{(r)} \tag{16}$$

Taking for all objects and all subsets a coefficient b_{ij} and the weight associated to x_{il} is similar to taking separately each possible weight $w_{kl}^{(r)}$ of the attribute $\mathbf{A}_l$ and summing the coefficients b_{ij} associated to objects having the same value $a_l^{(r)}$ and subsets containing ω_k. This leads to write J'_{ECM} as follow:

$$J'_{ECM}(w_{kl}) = \sum_{t=1}^{n_l} w_{kl}^{(t)} \underbrace{\sum_{A_j \subseteq \Omega, \omega_k \in A_j} \sum_{i \in \{1,\ldots,n\}/x_{il}=a_l^{(t)}} b_{ij}}_{cst}$$

The two last sums correspond to a constant. Thus, the maximization of the objective function under constraint (9) is a linear optimization problem with linear constraints. Optimal solution is given by eq. (14).

□

As can be observed, the minimization of the weight usually provides a crisp centroid for singletons. For other subsets with higher cardinality, fuzzy centroids are expected to appear more often, since they are defined as an average of singletons weights.

The algorithm of our proposed method is summarized in **Algorithm 1**. During initialization, singletons weights are randomly fixed such that Eq. (3) is satisfied for all attributes and clusters. Then, cluster centers with cardinality ≥ 2 are computed. Convergence is reached when centroids do not change from one iteration to another.

Algorithm 1 cat-ECM algorithm

Require: $\mathbf{X} = \{x_1, \ldots, x_n\}$ the categorical data, $1 < c < n$ the number of clusters, $\alpha \geq 1$ the weighting exponent for cardinality, $\beta > 1$ weighting exponent, and $\delta > 0$ the distance to the empty set.
Randomly initialize $\mathbf{W}$ that respect (3) and (9)
$t \leftarrow 0$
repeat
$t \leftarrow t + 1$
Update M using (12) and (13)
Update centroids V_{t-1} using (14)
until $V_{t-1} = V_t$

IV. Experimental results

A. Methodology

In order to validate cat-ECM, three categorical data sets were used: Cat-diamond, a categorical toy data set inspired from the diamond data set [10] (cf. Fig. 1), Soybean and Zoo, two classical data sets available on the UCI Machine Learning repository.Characteristics of each data set are reported in Table I. It should be emphasized that the data sets contain categorical attributes with no order properties on the values. Fig. 1 shows one possible representation of the data points for Cat-diamond; others are possible and equally valid.

TABLE I: Characteristics of the datasets.

	objects number	attributes number	classes number
Cat-diamond	12	2	2
Soybean	47	35	4
Zoo	101	17	7

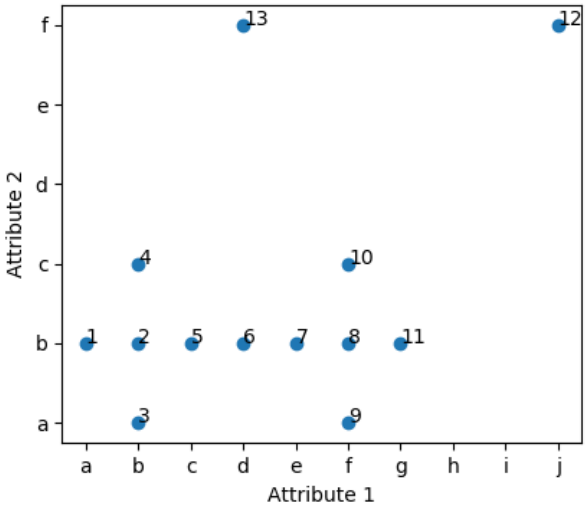

Fig. 1: Cat-diamond data set

Since real classes are known for these data sets, the performance of cat-ECM was assessed using two classical measures: (1) the adjusted rand index (ARI) [20], which computes a similarity measure between two crisp partitions, and (2) the Huang's accuracy [21], which directly compares the true classes with the crisp partition found. In order to obtain a crisp partition from our algorithm, a pignistic transformation is performed on the credal partition and then a maximal probabilistic rule is used. Furthermore, the behavior of cat-ECM was also analyzed using the normalized non-specificity measures on the generated credal partition.

In order to avoid local minima due to a random initialization of the centroids, an experiment consisted of running cat-ECM 10 times and selecting the solution giving the minimum value of the objective function.

B. Behavior of cat-ECM

We first analyzed the Cat-diamond toy data set with the following parameters: $c = 2$, $\alpha = -0.05$, $\beta = 1.1$ and $\delta = 1.05$. Fig. 2 presents the masses obtained by cat-ECM plotted against the objects.

As can be observed, objects 2, 3 and 4 are grouped in the first cluster and objects 8, 9 and 10 form the second cluster. This can be explained by the fact that they share the same value for the most discriminant attribute, i.e attribute 1. Note

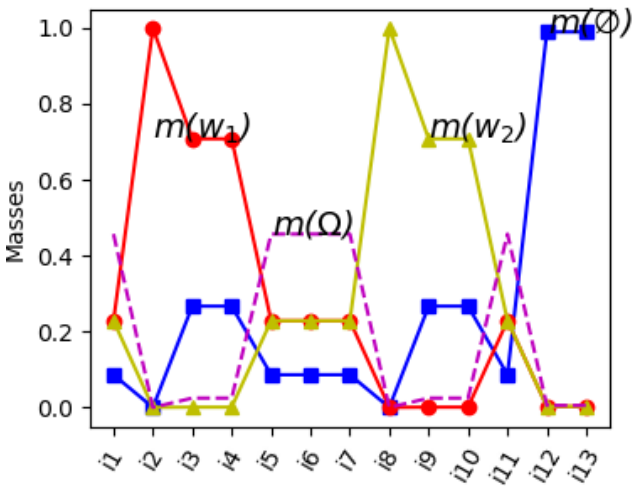

Fig. 2: Masses obtained with Cat-diamond data set

TABLE II: ARI, Accuracy and normalized non-specificity obtained on Soybean data set

α	ARI	Accuracy	N*(c)
-1	0.5	0.74	0.42
-0.05	0.82	0.94	0.05
0	0.87	0.96	0.03
1	0.87	0.96	0
2	1	1	0

that objects 2 and 8 are allocated with full certainty to cluster ω_1 and ω_2 respectively. Hence, the cluster centers are located on those two objects.

Object 2 in ω_1 and object 8 in ω_2 have in common with objects 1, 5, 6, 7, 11 the value of the attribute 2. Thus, $\mathbf{x}_1$, $\mathbf{x}_5$, $\mathbf{x}_6$, $\mathbf{x}_7$ and $\mathbf{x}_{11}$ belong to Ω, meaning that they are between the two clusters.

Finally, objects 12 and 13, with their allocation to the empty set, are considered as outliers. They are indeed far from all other objects.

C. Guidelines for parameters

Before running cat-ECM, parameters δ, β and α should be set. First, the δ value can be obtained using a rejection rate [10].

The β parameter controls the fuzziness of the credal partition. As in ECM, high values of β implies balanced masses for the clusters. We observed throughout the experiments that $\beta \geq 2$ corresponds to a high value for cat-ECM. Thus, we set $\beta = 1.1$ close to 1 for the rest of the experiments.

In order to test the α parameter, which controls the quantity of imprecision available in the final credal partition, we executed cat-ECM on the Soybean data set with various values of α. For this experiment, other parameters are fixed as follows: $\delta = 10$, $\beta = 1.1$, $c = 4$ and subsets are limited to the empty set and the ones with a cardinality ≤ 2. Results are reported Table II.

As can be observed with the ARI and accuracy results, our algorithm provides good performance on the Soybean data set, with higher values of α leading to better clustering solutions.

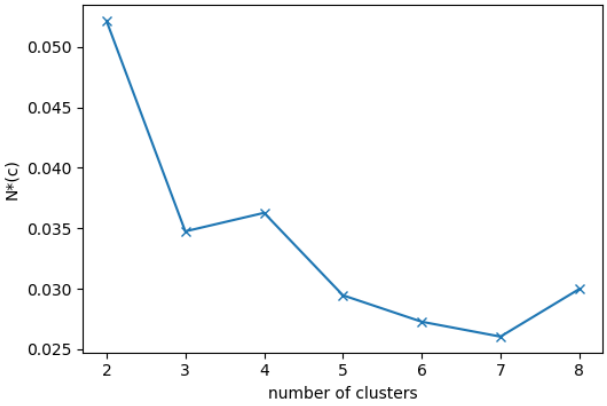

Fig. 3: Validity index on Zoo data set

Inversely, the normalized non-specificity is higher when α is low. For the next experiment, we decided to keep $\alpha = -0.05$, as this would allow the expression of uncertainties and would lead to reasonable clustering solutions.

It should be emphasized that in ECM, the α parameter is normally > 0 in order to penalize subsets with high cardinality. In cat-ECM, since the optimization of the prototypes provides hard values for the clusters, distances between subsets with a high cardinality and objects are usually further than distances between clusters and objects. Thus, to obtain allocation of belief to subsets with high cardinalities, singletons should be penalized by setting a negative value for α.

D. Validation of cat-ECM on Zoo data set

Usually in a concrete clustering problem, no background knowledge is available. Thus, the following experiment with the Zoo data set started with the assumption that the number of classes c was unknown.

In order to choose a relevant number of clusters, a classical method consists of measuring a validity index from partitions generated by the clustering algorithm with various values of c and analyzing the curve. In the framework of evidential clustering algorithms, a validity index frequently used is the normalized non-specificity measure. The minimum value of the measure corresponds to the experience with the optimal number of clusters [10].

The cat-ECM algorithm was performed with $\alpha = -0.05$, $\beta = 1.1$ and $\delta = 10$ and c varying from 2 to 8. The resultant normalized non-specificity values are presented Fig. 3.

As can be observed, the optimal number of clusters is reached for $c = 7$, which actually corresponds to the real number of classes. The credal partition obtained with $c = 7$ was transformed into hard credal partition by assigning each object to the subset of classes with the highest mass. The result of this is illustrated Fig. 4. A multiple components analysis was employed to plot data points in 2D. Note that the accuracy of the solution is 0.93.

Clusters C1 to C7 represent, respectively, reptiles, invertebrates, birds, amphibians, fishes, mammals and insects.

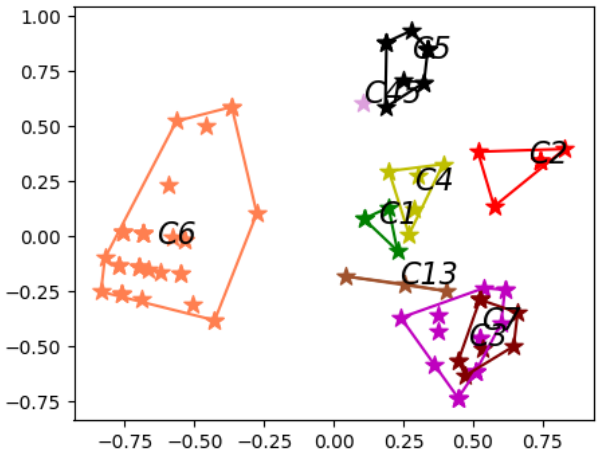

Fig. 4: cat-ECM with 7 clusters on Zoo data set

Objects in the subset C13={reptile,bird} correspond to the animals land tortoise, rhea and kiwi. The uncertainty between C1 and C3 can be explained as following: the land tortoise, which is in reality a reptile, is the only herbivorous reptile within the zoo data set. Thus, the distance with other reptiles is equivalent to the distance with herbivorous birds. The rhea and the kiwi are flightless birds; consequently, their assignment to the bird class is uncertain and reptiles are close to them. The object assigned to C45={amphibian,fish} corresponds to the sea-snake. It is in reality a reptile, but it can swim and its assignment to the subset C45 is therefore not surprising.

V. Conclusion

This paper presents a new categorical clustering algorithm referred to as cat-ECM. Similar to ECM, our method generates a credal partition which brings richer information about uncertainties and imprecision than a hard, a fuzzy or a possibilistic partition. The novelty of our approach corresponds to the introduction of weights for the centroids and the definition of a new dissimilarity measure between categorical objects and cluster centroids. An alternate minimization scheme is proposed to solve the clustering problem. While the update of the masses does not differ from ECM, the update of the centroids provides hard weighting coefficients. Preliminary results on three data sets show that cat-ECM is efficient for the analysis of data sets containing outliers and overlapping clusters. Additional validation work needs to be performed to understand how changes to the various parameters of cat-ECM affects the clustering solution, how these results vary with the number of objects in a data set, and how the performance of cat-ECM compares to closed categorical clustering methods. Nevertheless, the ability of cat-ECM to handle categorical data makes it highly useful for the analysis of survey data, which are common in for e.g. health research and which often contain categorical, discrete and continuous data types.

Acknowledgment

The authors acknowledge the support received from the Agence Nationale de la Recherche of the French government throught the program "Investissements d'Avenir"(16-IDEX-0001 CAP 20-25).

References

[1] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. Patel, A. Tiwari, M. Er, W. Ding, and C. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[2] M. M. Bamman, J. K. Petrella, J.-s. Kim, D. L. Mayhew, and J. M. Cross, "Cluster analysis tests the importance of myogenic gene expression during myofiber hypertrophy in humans," *Journal of Applied Physiology*, 2017.

[3] A. Moschetti, L. Fiorini, D. Esposito, P. Dario, and F. Cavallo, "Daily activity recognition with inertial ring and bracelet: An unsupervised approach," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pp. 3250–3255, IEEE, 2017.

[4] N. Lei and S. Moon, "A decision support system for market-driven product positioning and design," *Decision Support Systems*, vol. 69, pp. 82–91, 2015.

[5] A. Russo, C. Gouveia, R. Trigo, M. Liberato, and C. DaCamara, "The influence of circulation weather patterns at different spatial scales on drought variability in the iberian peninsula," *Frontiers in Environmental Science*, vol. 3, p. 1, 2015.

[6] J. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2, pp. 191 – 203, 1984.

[7] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," *IEEE transactions on fuzzy systems*, vol. 1, no. 2, pp. 98–110, 1993.

[8] H. Yu and J. Fan, "Cutset-type possibilistic c-means clustering algorithm," *Applied Soft Computing*, vol. 64, pp. 401–422, 2018.

[9] R. N. Dave, "Robust fuzzy clustering algorithms," in *Second IEEE International Conference on Fuzzy Systems*, pp. 1281–1286, IEEE, 1993.

[10] M. Masson and T. Denœux, "ECM: An evidential version of the fuzzy c-means algorithm," *Pattern Recognition*, vol. 41, no. 4, pp. 1384–1397, 2008.

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

[12] T. Denœux and O. Kanjanatarakul, "Evidential clustering: A review," in *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, pp. 24–35, Springer, 2016.

[13] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[14] Z. Huang and M. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 4, pp. 446–452, 1999.

[15] D. Kim, K. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern recognition letters*, vol. 25, no. 11, pp. 1263–1271, 2004.

[16] Z. He, X. Xu, and S. Deng, "Attribute value weighting in k-modes clustering," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15365–15369, 2011.

[17] H. Jia, Y. Cheung, and J. Liu, "A new distance metric for unsupervised learning of categorical data," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 5, pp. 1065–1079, 2016.

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

[19] G. Klir and M. Wierman, *Uncertainty-based information: elements of generalized information theory*, vol. 15. Physica, 2013.

[20] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[21] Z. Huang and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 4, pp. 446–452, 1999.