Categorical fuzzy entropy c-means

Abdoul Jalil Djiberou Mahamadou*, Violaine Antoine*, Engelbert Mephu Nguifo*, and Sylvain Moreno†
*Clermont Auvergne University, CNRS, ENSMSE, LIMOS, F-63000 Clermont-Ferrand, France
Email: {abdoul_jalil.djiberou_mahamadou, violaine.antoine, engelbert.mephu_nguifo}@uca.fr
†Digital Health Hub, Simon Fraser University, Vancouver, Canada
Email: sylvain_moreno@sfu.ca

***Abstract*—Hard and fuzzy clustering algorithms are part of the partition-based clustering family. They are widely used in real-world applications to cluster numerical and categorical data. While in hard clustering an object is assigned to a cluster with certainty, in fuzzy clustering an object can be assigned to different clusters given a membership degree. For both types of method an entropy can be incorporated into the objective function, mostly to avoid solutions raising too much uncertainties. In this paper, we present an extension of a fuzzy clustering method for categorical data using fuzzy centroids. The new algorithm, referred to as Categorical Fuzzy Entropy (CFE), integrates an entropy term in the objective function. This allows a better fuzzification of the cluster prototypes. Experiments on ten real-world data sets and statistical comparisons show that the new method can efficiently handle categorical data.**

***Index Terms*—clustering, fuzzy C-means, categorical data, entropy, fuzzy centroids**

I. Introduction

Clustering is a popular unsupervised learning method that aims at grouping data objects such that similar data belong to the same group (cluster) and dissimilar data to different groups. Clustering methods are generally categorized in three families: hierarchical, density-based and partition-based. In hierarchical clustering data objects are clustered using hierarchy of clusters by either agglomerative (bottom-up) or divisive (top-down) strategies. Density-based clustering are spatial clustering methods that group data given the most dense regions. In partition-based clustering methods a cluster membership partition which can be either hard or fuzzy is generated. The two well-known hard and fuzzy clustering algorithms are the k-means [1] and the fuzzy C-means (FCM) [2]. Although the k-means algorithm can efficiently cluster data, it is limited when expressing uncertainty of cluster assignment. Inversely, FCM is able to express uncertainties on the class memberships. Such possibility is relevant in many applications such as digital security [3], image segmentation [4], economy [5], agriculture [6], etc.

Entropy-based fuzzy clustering methods are extensions of the fuzzy clustering in which a weighted entropy is incorporated into the objective function. Depending on the application the entropy can have different roles and meanings. In [7] the entropy is seen as a regularizing function to the objective function of FCM. In [8] the authors used the entropy as a prior in Bayesian context for image restoration and proposed a new clustering method based on the fuzzy framework. In [9] an entropy-based fuzzy clustering method that automatically identifies the number and initial locations of cluster centers is proposed. In [10] the entropy of the membership functions is incorporated into the objective function to allow gradual transition from a maximum uncertainty to a minimum uncertainty during the clustering process. When applied for clustering validation, the entropy corresponds to an internal validity index [11]. In this case the entropy measures the fuzziness of partitions produced by clusters. More generally the entropy H of a probability p is defined as $H(p) = -p\log(p)$. The entropy H, also called the Shannon's entropy [12], has the following properties: it has a low value (respectively a high value) when the probability p is close to 0 (respectively when p is uniformly distributed).

To overcome the limitation of many clustering algorithms to numeric-only data, Huang initially proposed an adaptation of the k-means algorithm for categorical data [13]. In this method, named k-modes, a simple matching of attributes values is used as dissimilarity measure and the clusters prototypes are represented by the most frequent values (modes) of each attribute. Later, Huang proposed a generalization of the k-modes algorithm called fuzzy k-modes (FKM) [14]. In this method, data objects have a membership degree for each cluster. Various extensions of the FKM have been later proposed [15], [16], [17], [18], [19], [20], [21]. One limitation of FKM method is the misrepresentation of clusters when the frequencies are similar. This limit is related to the dependency of cluster prototypes on the attributes value frequencies. To address this problem the FKM algorithm has been extended to consider fuzzy centroids [21]. In the later method, rather than representing cluster prototypes by the most frequent values for each attribute, all categorical values of each attribute is associated to prototypes with a weight that is updated at each iteration.

However, it can be noticed that the updating formulas of cluster prototypes in [21] do no guarantee the convergence of the method. Therefore it can lead to non optimal solutions. In this paper we proposed new updating formulas of cluster prototypes presented in [21] and extend the method to entropy-based fuzzy clustering to have fuzzy representation of cluster prototypes.

The remaining paper is organized as follows: Section II introduces the fuzzy k-modes and the fuzzy centroids clustering algorithms. Section III describes the new updating formulas of cluster prototypes for the fuzzy centroids clustering and Section IV details the new entropy-based clustering algorithm. In the next section, Section V, the methodology of the experiences and the results are presented. It is finally followed by a conclusion and some perspective of the work in Section VI.

II. Clustering preliminaries

In this section we present the fuzzy k-modes and fuzzy centroids clustering methods.

A. Fuzzy k-modes

The fuzzy k-modes (FKM) [13] clustering algorithm is an extended version of the k-modes algorithm to cluster categorical data based on the fuzzy clustering framework. Let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be a collection of n categorical objects described by categorical attributes $A_1, A_2, \dots, A_p$. For the attribute A_l such that $1 \leq l \leq p$, there exists in its domain n_l possible categorical values: $DOM(A_l) = \{a_l^{(1)}, \dots, a_l^{(n_l)}\}$. Hence $\mathbf{x}_i = [x_{i1}, \dots, x_{il}, \dots, x_{ip}]$ is a vector of p observed features for the i^{th} object and x_{il} denotes the value of the l^{th} feature for the object $\mathbf{x}_i$. Let k be the number of clusters and $\mathbf{v}_j$ be the cluster center such that $\mathbf{v}_j = [v_{j1}, \dots, v_{jp}]$ for $1 \leq j \leq k$. The FKM objective function is given by:

$$J_{FKM}(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m d(\mathbf{x}_i, \mathbf{v}_j) \tag{1}$$

subject to

$$0 \leq u_{ij} \leq 1, \quad \forall 1 \leq i \leq n, 1 \leq j \leq k, \tag{2}$$

$$\sum_{j=1}^{k} u_{ij} = 1, \quad \forall 1 \leq i \leq n, \tag{3}$$

and

$$0 < \sum_{i=1}^{n} u_{ij} < n, \quad 1 \leq j \leq k, \tag{4}$$

where $\mathbf{U} = [u_{ij}]$ is the fuzzy partition matrix, m a coefficient controlling the fuzziness of the partition and $d(\mathbf{x_i}, \mathbf{v}_j)$ is the dissimilarity measure given by

$$d(\mathbf{x_i}, \mathbf{v}_j) = \sum_{l=1}^{p} \delta(x_{il}, v_{jl}), \tag{5}$$

where

$$\delta(x_{il}, v_{jl}) = \begin{cases} 0 & if \quad x_{il} = v_{jl}, \\ 1 & x_{il} \neq v_{jl}. \end{cases}$$

The optimization problem can be solved by a partial optimization scheme which consists in fixing the variable $\mathbf{U}$ and solving the reduced problem given $\mathbf{V}$ and inversely. The process is iterated until convergence. The updating formulas of $\mathbf{U}$ and $\mathbf{V}$ are given by:

$$u_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i = \mathbf{v}_j, \\ 0 & \text{if } \mathbf{x}_i = \mathbf{v}_h, h \neq j, \\ 1 \Big/ \sum_{h=1}^{k} \left[\dfrac{d(\mathbf{x}_i, \mathbf{v}_j)}{d(\mathbf{x}_i, \mathbf{v}_h)} \right]^{\frac{1}{m-1}} & \text{if } \mathbf{x}_i \neq \mathbf{v}_h, 1 \leq h \leq k, \end{cases} \tag{6}$$

and $v_{jl} = a_l^{(r)} \in DOM(A_l)$ where

$$\sum_{i, x_{ij} = a_l^{(r)}} u_{ij}^m \geq \sum_{i, x_{ij} = a_l^{(t)}} u_{ij}^m, \quad \forall 1 \leq t \leq n_l, t \neq r. \tag{7}$$

As it can be observable, values of cluster prototypes correspond to the most frequent ones. Such crisp decision on the centroids can lead to the distortion of the cluster representation. Let us define for instance an attribute a_l with $DOM(A_l) = \{a_l^{(1)}, a_l^{(2)}\}$, and let us consider that the frequencies of this domain values are respectively 20 and 19. In this case the prototype corresponding to A_l is $a_l^{(1)}$ despite uncertainties due to the closed frequencies of $a_l^{(1)}$ and $a_l^{(2)}$ can better represent the centroid. To overcome this problem a generalization of the FKM using fuzzy centroids has been proposed in [21].

B. Fuzzy centroids clustering

The fuzzy centroids (FC) clustering [21] considers fuzzy centroid with fuzzy categorical values, the objective function is similar to the FKM's. The difference lies in the prototypes definition.

Let $\mathbf{v}_j = (v_{j1}, \dots, v_{jl}, \dots, v_{jp})$ be a the prototype of cluster j. In FKM, v_{jl} represents the mode of attribute l whereas it is defined in FC as follows:

$$v_{jl} = [w_{jl}^{(1)} a_l^{(1)} \wedge \dots \wedge w_{jl}^{(n_l)} a_l^{(n_l)}] \tag{8}$$

subject to

$$0 \leq w_{jl}^{(t)} \leq 1, \quad 1 \leq t \leq n_l \tag{9}$$

$$\sum_{t=1}^{n_l} w_{jl}^{(t)} = 1, \quad 1 \leq l \leq p. \tag{10}$$

and the dissimilarity measure by

$$d(\mathbf{x_i}, \mathbf{v}_j) = \sum_{l=1}^{p} \sum_{t=1}^{n_l} \delta(x_{il}, a_l^{(t)}), \tag{11}$$

where

$$\delta(x_{il}, a_l^{(t)}) = \begin{cases} 0 & \text{if} \quad x_{il} = a_l^{(t)}, \\ w_{jl}^{(t)} & \text{if} \quad x_{il} \neq a_l^{(t)}. \end{cases}$$

In (8), $w_{jl}^{(t)}$ corresponds to the weight associated to cluster j, attribute l, and the t^{th} categorical value. With the later definition the prototypes do no longer depend on the frequency of the attributes values but consider the combination of all values.

Similarly to the FKM, the optimization problem can be solved by a partial optimization. Hence, when $\mathbf{V}$ is fixed the updating formulas of u_{ij} is given by equation (6). Next, when $\mathbf{U}$ is fixed the updating formulas of $w_{jl}^{(t)}$ given in [21] are

$$w_{jl}^{(t)} = \sum_{i=1}^{n} \gamma(x_{il}), \tag{12}$$

where

$$\gamma(x_{il}) = \begin{cases} u_{ij}^m & \text{if} \quad x_{il} = a_l^{(t)}, \\ 0 & \text{if} \quad x_{il} \neq a_l^{(t)}. \end{cases}$$

III. New fuzzy centroids updating formulas

We carried out experiences on different data sets and parameters setting of the original fuzzy centroids algorithm and observed in some experiences the cost function is not monotonically decreasing. In this section we rigorously derive the objective function of the fuzzy centroids method and showed that the updating formulas (12) do not guaranty the convergence of the method.

Theorem 1. *Let $S_{il}^{(t)}$ defined by*

$$S_{il}^{(t)} = \sum_{i,x_{ij}=a_l^{(t)}} u_{ij}^m .$$

For $\mathbf{U}$ *fixed the objective function* (1) *is minimized iff*

$$w_{jl}^{(r)} = \begin{cases} 1 & if\ S_{il}^{(r)} > S_{il}^{(t)}, \forall t \in \{1,\ldots,n_l\}, t \neq r, \\ 0 & if\ S_{il}^{(r)} < S_{il}^{(t)}, \exists t \in \{1,\ldots,n_l\}, t \neq r \\ 1 - \sum_{\tau=1}^{q-1} w_{jl}^{\tau} & if\ S_{il}^{(r)} = S_{il}^{(\tau_1)} = \ldots = S_{il}^{(\tau_{q-1})} > S_{il}^{(t)}, \\ & \forall \tau_1,\ldots,\tau_{q-1}, t \in \{1,\ldots,n_l\}\ s.t. \\ & \tau_1 \neq \ldots \neq \tau_{q-1} \neq t. \end{cases} \tag{13}$$

Proof: The objective function (1) with distance (11) can be rewritten as

$$J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^m \sum_{l=1}^{p} \sum_{t,a_l^{(t)} \neq x_{il}} w_{jl}^{(t)} .$$

From (10) we have

$$\sum_{t,a_l^{(t)} \neq x_{il}} w_{jl}^{(t)} = 1 - \sum_{t,a_l^{(t)} = x_{il}} w_{jl}^{(t)} . \tag{14}$$

Since the sums on objects, clusters and dimensions are independent and using (14) we can rewrite the objective function as

$$J(\mathbf{U},\mathbf{V}) = \sum_{j=1}^{k}\sum_{l=1}^{p}\sum_{i=1}^{n} \big[u_{ij}^m - u_{ij}^m \sum_{t,a_l^{(t)}=x_{il}} w_{jl}^{(t)}\big].$$

Minimizing J is equivalent to minimizing J_1 $\forall l \in [1,p]$ where

$$J_1(\mathbf{U}_{.j},\mathbf{V}_{.j}) = \sum_{i=1}^{n} u_{ij}^m - \sum_{i=1}^{n} u_{ij}^m \sum_{a_l^{(t)}=x_{il}} w_{jl}^{(t)} .$$

Since $\mathbf{U}$ is fixed, minimizing J_1 is equivalent to maximizing

$$J_2(\mathbf{V}_{.j}) = \sum_{i=1}^{n} u_{ij}^m \sum_{t,a_l^{(t)}=x_{il}} w_{jl}^{(t)},$$

under the constraints (9) and (10). The weights w_{jl} are independent hence the maximization of J_2 can be rewritten as the maximization of J_3 $\forall j \in [1,k]$:

$$J_3(w_{jl}) = \sum_{i=1}^{n} \sum_{t,a_l^{(t)}=x_{il}} u_{ij}^m w_{jl}^{(t)}$$

Since

$$\sum_{i=1}^{n} \sum_{t,a_l^{(t)}=x_{il}} u_{ij}^m w_{jl}^{(t)} = \sum_{t=1}^{n_l} \sum_{i,x_{il}=a_l^{(t)}} u_{ij}^m w_{jl}^{(t)} \tag{15}$$

then the optimization problem becomes

$$\begin{cases} \max & J_2(w_{jl}) = \sum_{t=1}^{n_l} \sum_{i,x_{il}=a_l^{(t)}} u_{ij}^m w_{jl}^{(t)}, \\ s.t. & \sum_{t=1}^{n_l} w_{jl}^{(t)} = 1. \end{cases} \tag{16}$$

For $\mathbf{U}$ fixed, the term $\sum_{i,x_{il}=a_l^{(t)}} u_{ij}^m$ is constant hence J_2 is maximized if only and if $w_{jl}^{(t)}$ satisfies the equation (13). ■

Equation (13) shows that the updating formulas proposed in [21] do not guaranty the convergence of the fuzzy centroids algorithm. In the remaining paper we denote FC* the fuzzy centroids clustering with correct updating formulas of the prototypes.

One can note that the new update of $w_{jl}^{(t)}$ gives most of the time binary values. Indeed in pratice the case $S_{il}^{(r)} = S_{il}^{(\tau_1)} = \ldots = S_{il}^{(\tau_{q-1})}$ is very unlikely to appear. Hence the algorithm generates mostly hard centroids instead of fuzzy. To overcome this problem we propose an entropy-based fuzzy centroids method.

IV. Entropy-based fuzzy c-means with fuzzy centroids

A. Objective function

The entropy-based fuzzy centroids method, called Categorical Fuzzy Entropy (CFE), is a variant of FC* that incorporates an entropy penalization term in the objective function. Such term allows a trade-off between hard and fuzzy centroids: the update formulas for the weights do not lead to binary values and the entropy term penalize uniform weights leading to total uncertainty. We define the objective function of the CFE as follow:

$$J_{CFE}(\mathbf{U},\mathbf{V}) = J_{FKM}(\mathbf{U},\mathbf{V}) + \alpha n \sum_{j=1}^{k}\sum_{l=1}^{p}\sum_{t=1}^{n_l} w_{jl}^{(t)} log(w_{jl}^{(t)}) \tag{17}$$

under the constraints (2), (3), (4), (9) and (10). The α parameter is a weighted coefficient that controls the importance of the within cluster criteria and the entropy.

B. Optimization

Similarly to the FKM and FC, J_{CFE} can be solved using an alternate optimization scheme. First we consider that $\mathbf{V}$ is fixed. In that case, the constraint minimization problem with respect to $\mathbf{U}$ is identical to FKM and the solution is given by equation (6). Second, we consider $\mathbf{U}$ fixed and obtain the following theorem.

Theorem 2. *For* $\mathbf{U}$ *fixed, the cluster prototypes* $\mathbf{V}$ *are minimized iff*

$$w_{jl}^{(t)} = \frac{\exp\Big[-\frac{1}{n\alpha}\sum_{x_{il}\neq a_l^{(t)}} u_{ij}^m\Big]}{\sum_{t=1}^{n_l}\exp\Big[-\frac{1}{n\alpha}\sum_{x_{il}\neq a_l^{(t)}} u_{ij}^m\Big]} . \tag{18}$$

Proof: We can rewrite J_{CFE} as

$$J_{CFE}(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\Big[\sum_{l=1}^{p}\sum_{t,a_l^{(t)}\neq x_{il}} u_{ij}^{m} w_{jl}^{(t)} + \alpha\sum_{l=1}^{p}\sum_{t=1}^{n_l} w_{jl}^{(t)} log(w_{jl}^{(t)})\Big]$$

Since $\mathbf{U}$ is fixed and the sum over clusters and attributes are independent, given $j \in [1,k]$ and $l \in [1,p]$, optimizing J_{CFE} is equivalent to optimizing

$$J_4(w_{jl}) = \sum_{i=1}^{n}\sum_{t,a_l^{(t)}\neq x_{il}} u_{ij}^{m} w_{jl}^{(t)} + \alpha n\sum_{t=1}^{n_l} w_{jl}^{(t)} log(w_{jl}^{(t)})$$

Using (15), minimizing the new objective function is equivalent to minimizing

$$\begin{cases} J_5(w_{jl}) = \displaystyle\sum_{t=1}^{n_l}\sum_{i,x_{il}\neq a_l^{(t)}} u_{ij}^{m} w_{jl}^{(t)} + \alpha n\sum_{t=1}^{n_l} w_{jl}^{(t)}\log(w_{jl}^{(t)}), \\ s.t. \displaystyle\sum_{t=1}^{n_l} w_{jl}^{(t)} = 1. \end{cases}$$

Let $\mathcal{L} = J_5(w_{jl}) + \lambda_{jl}(\sum_{t=1}^{n_l} w_{jl}^{(t)} - 1)$ be the Lagrangian associated to the optimization problem where λ_{jl} is a Lagrangian multiplier. By differentiating the Lagrangian with respect to $w_{jl}^{(s)}$ and λ_{jl} we obtain:

$$\frac{\partial\mathcal{L}}{\partial w_{jl}^{(s)}} = \Big[\sum_{i,x_{il}\neq a_l^{(s)}} u_{ij}^{m}\Big] + \alpha n(1 + \log(w_{jl}^{(s)})) + \lambda_{jl}, \tag{19}$$

$$\frac{\partial\mathcal{L}}{\partial \lambda_{jl}} = \sum_{t=1}^{n_l} w_{jl}^{(t)} - 1. \tag{20}$$

Setting equation (19) to 0 gives

$$w_{jl}^{(s)} = \exp\Big[-\Big(1 + \frac{\lambda_{jl}}{n\alpha} + \frac{1}{n\alpha}\sum_{i,x_{il}\neq a_l^{(s)}} u_{ij}^{m}\Big)\Big]. \tag{21}$$

Setting equation (20) to 0 and replacing $w_{jl}^{(s)}$ by equation (21) gives

$$\exp\Big[-\Big(1 + \frac{\lambda_{jl}}{n\alpha}\Big)\Big] = \frac{1}{\displaystyle\sum_{t=1}^{n_l}\exp\Big[-\frac{1}{n\alpha}\sum_{i,x_{il}\neq a_l^{(t)}} u_{ij}^{m}\Big]}. \tag{22}$$

reporting $\exp\Big[-\Big(1 + \frac{\lambda_{jl}}{n\alpha}\Big)\Big]$ into the equation (21) gives

$$w_{jl}^{(t)} = \frac{\exp\Big[-\frac{1}{n\alpha}\sum_{i,x_{il}\neq a_l^{(t)}} u_{ij}^{m}\Big]}{\sum_{t=1}^{n_l}\exp\Big[-\frac{1}{n\alpha}\sum_{i,x_{il}\neq a_l^{(t)}} u_{ij}^{m}\Big]}. \tag{23}$$

Hence J_{CFE} is minimized iff $w_{jl}^{(t)}$ satisfies equation (23). ■

The algorithm of our proposed method is summarized in **Algorithm 1**. Given the number of clusters c, a chosen value of m and α, the first step consists in initializing the centroids such that equations (9) and (10) are satisfied. Then the cluster membership degrees u_{ij} and the prototypes are updated using respectively equations (6) and (23). The preceding step is repeated until there exists almost no change from an iteration to another, i.e when $||V_{t-1} - V_t||$ reaches a variable ε set to a small value.

Algorithm 1 CFE algorithm

```
Require: $\mathbf{X} = \{x_1, \ldots, x_n\}$ the categorical data, $1 < c < n$
    the number of clusters, $\alpha > 0$ the fuzzy entropy weighting
    coefficient, $m > 1$ weighting exponent, and $\epsilon$ a stop criteria.
Begin
    Randomly initialize $\mathbf{V}_0$ that respects (9) and (10).
    $t \leftarrow 0$
    repeat
        $t \leftarrow t + 1$
        Update $U_t$ using (6)
        Update centroids $V_t$ using (18)
    until $||V_{t-1} - V_t|| < \varepsilon$
End
```

V. Experimental results

A. Methodology

In order to validate the proposed method we used ten categorical and real-word data sets of different size available on the UCI Machine Learning repository [22]: Zoo, Soybean , Congressional voting records, Breast Cancer, Lung, Cars, Mushrooms, Credits, Dermatology, and Connect-4. For each data set we compare the performance of the proposed method against the FKM and FC*. Characteristics of the data sets are detailed in Table I.

TABLE I: Categorical data sets

	# Objects	# Attributes	# Classes
Lung	32	56	3
Soybean	40	55	4
Zoo	101	41	7
Breast Cancer	286	9	2
Dermatology	366	34	5
Votes	434	16	2
Credits	689	15	2
Cars	1728	6	4
Mushrooms	8124	22	7
Connect-4	67576	42	3

We carried out 100 trials with different centroids initialization and set the parameter α to 0.01 since it gives efficient performances on the ten data sets. As a matter of fact, through experiments we noticed that larger values of the fuzzy entropy coefficient α leads to uniform weights.

For each trial we executed several times the algorithms using the fuzziness coefficient m between 1.1 and 2. In order to obtain fair comparisons, methods are using the same centroids initializations.

Since the real classes are known for the data sets, we used the Rand Index (RI) [23] to evaluate the performance of the methods. Considering this criteria, the method giving the best partition corresponds to the highest RI. Results are detailed next section.

B. Cost function comparison on fuzzy centroids clusterings

In the first part of the experiences, the objective function values for FC and FC* are compared. Figure 1 presents the cost over 100 iterations with different initialization on the Zoo data set for the FC and FC* methods.

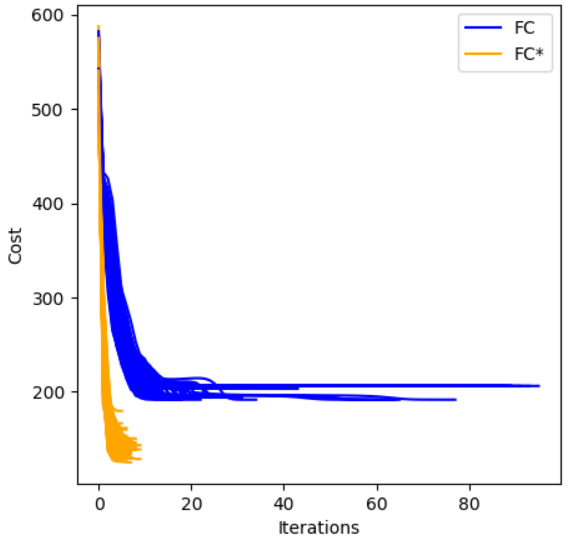

Fig. 1: Cost of FC and FC* over 100 iterations on zoo data set

As it can be seen, for all the iterations the costs for the FC* method are below the FC ones. Similarly, the number of iterations necessary to converge is lower for the FC* method. These observations can be explained by the fact the correct updating formulas of $w_{jl}^{(t)}$ help the objective function to converge faster and better.

Note that the behavior of non monotonically decrease of the objective function for FC is not visible through figure 1. However it has been numerically observed with some initializations.

C. Accuracy comparison

We compared the FKM and the original fuzzy centroids algorithm with the correct prototypes updating formulas denoted by FC*, and our entropy-based method CFE.

First, Figure 2 presents the average rand index with error bars of FC* and CFE for various values of m for the soybean data set.

As it can be noticed, for the Soybean data set the proposed method performs better and has lower variance. These results can be interpreted as follows: the fuzzy centroids obtained using the entropy help to better represent clusters by considering several attributes values. Hence it allows to capture more information about the data.

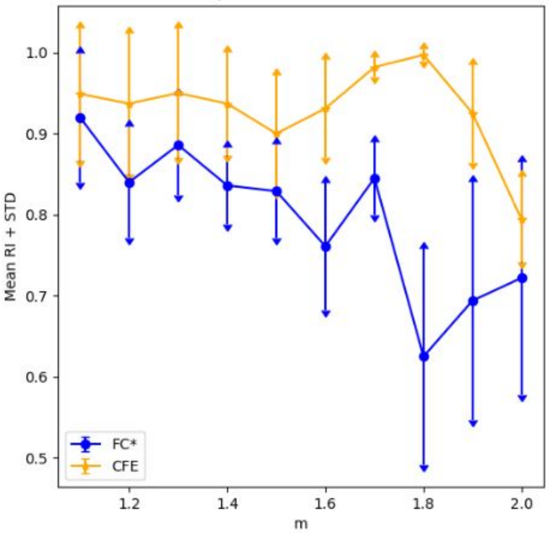

Fig. 2: Average RI and standard deviation varying with m for Soybean data set

We observed that depending on the data sets, best results for CFE are obtained with various m. Therefore this parameter has to be carefully set.

In order to statistically evaluate differences between the methods, a Friedman rank test [24] at significance level 0.05 was carried out on the average RI over the 100 trials. For each coefficient m, we computed the resulting critical difference (CD) diagram of Nemenyi post-hoc tests [24]. Due to space consideration only the CD of $m = 1.2$ is shown in Figure 3. We observed similar results for other values of m. Materials for computing the CD are from [25].

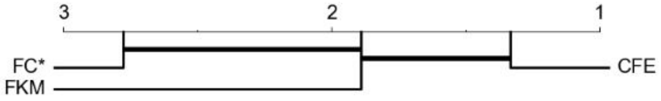

Fig. 3: Critical Diagram difference for m=1.2 where bold lines indicate groups of methods which are not significantly different.

From Figure 3, we can note that CFE is ranked 1 in the critical diagram difference. The CD shows that over the 100 trials, the CFE performs better than FKM and FC*. While the performance of CFE is significantly different from FC*, the difference is not significant between CFE and FKM and between FC* and FKM.

D. Discussions

We showed using different data sets that the proposed method can efficiently handle categorical data. However, our new method is not addressing or proposing solutions to two classic issues in the literature: defining the values of parameters m and α. The choice of the m value is a well known issue in fuzzy clustering [26] and there is little theoretical guidance in the literature. In general, the value of m is determined through experiences. Similarly to m, there is no general approach to determine the optimal value of α. We observed in the experiences that the α parameter behave as m when it is set with a high value. Indeed, it is known that if the value of m increases, the membership degrees u_{ij} become uniformly distributed: $u_{ij} = 1/k, \forall 1 \leq j \leq k$. In our case, when the value of α increases, the attributes values weights $w_{jl}^{(t)}$ becomes uniformly distributed: $w_{jl}^{(t)} = 1/n_l$, $\forall 1 \leq t \leq n_l$. Therefore, all the clusters prototypes become coincident, which harms the performance of the method.

VI. Conclusion

The k-modes algorithm is an adaptation of the k-means algorithm that handles categorical data. It has then been extended to the fuzzy framework by [14], who created a fuzzy centroids clustering algorithm called FC. In this paper, we first show that the updating formulas for cluster prototypes of FC do not guarantee the convergence of the method. Then, we propose a new clustering algorithm for categorical data that incorporates an entropy penalization term into the objective function. The goal of our method, called Categorical Fuzzy Entropy (CFE), is to better define fuzzy cluster prototypes. The interest of CFE is validated using ten real-world and categorical data sets. It has been compared with the fuzzy k-modes and the original fuzzy centroids clustering with the correct updating formulas. The results of the experiments concerning the comparison between the original fuzzy centroids algorithm with the correct prototypes updating formulas, denoted by FC*, and CFE illustrates that the later outperforms FC*. Statistical comparisons of FKM, FC* and CFE, show that on average for all values of m between 1.1 and 2, CFE performs better on all the considered data sets of the experiments. Our new method can then be used to efficiently handle categorical data.

In the future, we intend to better study the α parameter in order to propose some guidelines for its setting. Next, we plan to extend the proposed method to handle both numeric and categorical data.

Acknowledgment

We sincerely thank the anonymous reviewers whose constructive comments helped to improve this manuscript.

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

References

[1] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

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

[3] N. Naik, P. Jenkins, N. Savage, and L. Yang, "Cyberthreat hunting-part 2: Tracking ransomware threat actors using fuzzy hashing and fuzzy c-means clustering," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 23-26 June 2019, New Orleans.*, 2019.

[4] D. Kumar, H. Verma, A. Mehra, and R. Agrawal, "A modified intuitionistic fuzzy c-means clustering approach to segment human brain mri image," *Multimedia Tools and Applications*, vol. 78, no. 10, pp. 12663–12687, 2019.

[5] M. J. Rezaee, M. Jozmaleki, and M. Valipour, "Integrating dynamic fuzzy c-means, data envelopment analysis and artificial neural network to online prediction performance of companies in stock exchange," *Physica A: Statistical Mechanics and its Applications*, vol. 489, pp. 78–93, 2018.

[6] B. Choubin, K. Solaimani, M. H. Roshan, and A. Malekian, "Watershed classification by remote sensing indices: a fuzzy c-means clustering approach," *Journal of Mountain Science*, vol. 14, no. 10, pp. 2053–2063, 2017.

[7] D. Tran and M. Wagner, "Fuzzy entropy clustering," in *Ninth IEEE International Conference on Fuzzy Systems. FUZZ- IEEE 2000*, vol. 1, pp. 152–157 vol.1, May 2000.

[8] A. Lorette, X. Descombes, and J. Zerubia, "Fully unsupervised fuzzy clustering with entropy criterion," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 3, pp. 986–989 vol.3, Sep. 2000.

[9] J. Yao, M. Dash, S. Tan, and H. Liu, "Entropy-based fuzzy clustering and fuzzy modeling," *Fuzzy Sets and Systems*, vol. 113, no. 3, pp. 381 – 388, 2000.

[10] N. B. Karayiannis, "Meca: maximum entropy clustering algorithm," in *Proceedings of 1994 IEEE 3rd International Fuzzy Systems Conference*, pp. 630–635 vol.1, June 1994.

[11] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms.* USA: Kluwer Academic Publishers, 1981.

[12] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[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] G. Gan, J. Wu, and Z. Yang, "A genetic fuzzy k-modes algorithm for clustering categorical data," *Expert Systems with Applications*, vol. 36, no. 2, Part 1, pp. 1615 – 1620, 2009.

[16] Chi-Hyon Oh, K. Honda, and H. Ichihashi, "Fuzzy clustering for categorical multivariate data," in *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*, vol. 4, pp. 2154–2159 vol.4, July 2001.

[17] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "Multiobjective genetic algorithm-based fuzzy clustering of categorical attributes," *IEEE Transactions on Evolutionary Computation*, vol. 13, pp. 991–1005, Oct 2009.

[18] L. Bai, J. Liang, C. Dang, and F. Cao, "A novel fuzzy clustering algorithm with between-cluster information for categorical data," *Fuzzy Sets and Systems*, vol. 215, pp. 55 – 73, 2013.

[19] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," *Knowledge-Based Systems*, vol. 30, pp. 129 – 135, 2012.

[20] T. Li and Y. Chen, "Fuzzy clustering ensemble algorithm for partitioning categorical data," in *2009 International Conference on Business Intelligence and Financial Engineering*, pp. 170–174, July.

[21] 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.

[22] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[23] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[24] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[25] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Mining and Knowledge Discovery*, vol. 33, no. 4, pp. 917–963, 2019.

[26] Jian Yu, Qiansheng Cheng, and Houkuan Huang, "Analysis of the weighting exponent in the FCM," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, pp. 634–639, Feb 2004.