applied sciences

MDPI

Article

A Maximum-Entropy Fuzzy Clustering Approach for Cancer Detection When Data Are Uncertain

Mario Fordellone [1,†], Ilaria De Benedictis [2,*,†], Dario Bruzzese [3] and Paolo Chiodini [1]

[1] Medical Statistics Unit, Universitiy of Campania "Luigi Vanvitelli", 81100 Naples, Italy
[2] Universitiy of Campania "Luigi Vanvitelli", 81100 Naples, Italy
[3] Department of Public Health, University of Naples Federico II, 80131 Naples, Italy
* Correspondence: ilaria.debenedictis@gmail.com
† These authors contributed equally to this work.

Abstract: (1) Background: Cancer is a leading cause of death worldwide and each year, approximately 400,000 children develop cancer. Early detection of cancer greatly increases the chances for successful treatment, while screening aims to identify individuals with findings suggestive of specific cancer or pre-cancer before they have developed symptoms. Precise detection, however, often mainly relies on human experience and this could suffer from human error and error with a visual inspection. (2) Methods: The research of statistical approaches to analyze the complex structure of data is increasing. In this work, an entropy-based fuzzy clustering technique for interval-valued data (EFC-ID) for cancer detection is suggested. (3) Results: The application on the Breast dataset shows that EFC-ID performs better than the conventional FKM in terms of AUC value (EFC-ID = 0.96, FKM = 0.88), sensitivity (EFC-ID = 0.90, FKM = 0.64), and specificity (EFC-ID = 0.93, FKM = 0.92). Furthermore, the application on the Multiple Myeloma data shows that EFC-ID performs better than the conventional FKM in terms of Chi-squared (EFC-ID = 91.64, FKM = 88.26), Accuracy rate (EFC-ID = 0.71, FKM = 0.60), and Adjusted Rand Index (EFC-ID = 0.33, FKM = 0.21). (4) Conclusions: In all cases, the proposed approach has shown good performance in identifying the natural partition and the advantages of the use of EFC-ID have been detailed illustrated.

Keywords: cancer detection; cancer classification; unsupervised classification; entropy regularization procedure; penalized classification model; interval-valued data; imprecise data

Citation: Fordellone, M.; De Benedictis, I.; Bruzzese, D.; Chiodini, P. A Maximum-Entropy Fuzzy Clustering Approach for Cancer Detection When Data Are Uncertain. *Appl. Sci.* **2023**, *13*, 2191. https://doi.org/10.3390/app13042191

Academic Editors: Stefano Silvestri and Francesco Gargiulo

Received: 31 December 2022
Revised: 2 February 2023
Accepted: 6 February 2023
Published: 8 February 2023

1. Introduction

Cancer is a leading cause of death worldwide, accounting for nearly 10 million yearly deaths. Moreover, each year, approximately 400,000 children develop cancer. Cancer mortality is reduced when cases are detected and treated early. There are two components of early detection: early diagnosis and screening. Early detection of cancer greatly increases the chances for successful treatment, while screening aims to identify individuals with findings suggestive of specific cancer or pre-cancer before they have developed symptoms. Precise detection, however, often mainly relies on human experience and this could suffer from human error and error with a visual inspection. To try to solve these problems there is a demand for statistical/mathematical algorithms (e.g., supervised and unsupervised classification models, machine learning approaches, latent class analysis, etc.) for the early detection of tumors. Then, the efficiency and effectiveness of early diagnosis and screening can be increased if tumors are detected and classified automatically through computers [1].

In the conventional statistical data analysis, usually point data are analyzed, i.e., exact results of measurements that consist of features of the reference sample. These values can be either directly observed as results of measurements (e.g., systolic and/or diastolic blood pressure of a person) or can be observed as counts of a category (i.e., group) representing called events (e.g., the gender of that person). However, in many real life applications, the results of these measurements are never precise, and some degree of uncertainty that

characterizes them exists.

The uncertainty of a measurement can be defined as the interval on the measurement scale within which the true value lies with a specified probability when all sources of error have been taken into account [2]. The quantification of this uncertainty could become an important issue to treat in the area of the statistical quality of data. In the medical field, chemists/biologists should be expected as standard practice to provide a statement of the uncertainty alongside their estimated measure to make it into account in the data analysis step [2,3]. In other words, a measurement cannot be properly interpreted without knowledge of its uncertainty. In clinical practice, many rules and guidelines have been proposed aiming to provide a general overview of the uncertainty concept in the measurement step and make it into account for the data interpretation [4]. For some example, the reader can refer to [5] which provided a review where a rule-based approach is suggested with a number of the more common rules tabulated for the routine calculation of measurement uncertainty, and [6] which provided a systematic review regarding uncertainty tolerance in the health and healthcare-related outcomes.

The research of statistical approaches to analyze the complex structures of data is increasing. A lot of attention is focused on the methodologies to treat complex datasets where the data features are uncertain (called *imprecise data*). The simplest structure of *imprecise data* is the *interval-valued data* [7–9]. An interval-valued data can be formalized as $x_{ij} = [\underline{x}_{ij}, \bar{x}_{ij}]$, $i = 1, \ldots, n$ and $j = 1, \ldots, J$, where x_{ij} is the j-th interval-valued variable observed on the i-th observation, $\underline{x}_{ij}$ and $\bar{x}_{ij}$ denote the lower and upper bounds of the interval, respectively, (i.e., the extreme values registered for the j-th interval-valued variable on the i-th observation). Then, in a $n \times J$ interval-valued data matrix, each observation is represented as a hyperrectangle (in $\mathbb{R}^J$) having 2^J vertices [10]. However, in this work, we use a simpler notation of interval-valued data, consisting to consider centers and radii, separately. In particular *i* for the centers we indicate $\mathbf{C}$ the $n \times J$ *centers matrix* whose generic element $c_{ij} = 2^{-1}(\underline{x}_{ij} + \bar{x}_{ij})$ is the center (i.e., the midpoint) of the associated interval; *ii* for the radii we define $\mathbf{R}$ the $n \times J$ *radii matrix* whose generic element $r_{ij} = 2^{-1}(\bar{x}_{ij} - \underline{x}_{ij})$ is the radius of the associated interval. Then, by considering this reformulation of the interval-valued data, the complete interval-valued matrix can be formalized as follows:

$$\mathbf{X} \equiv \left\{ x_{ij} = [c_{ij}, r_{ij}] : i = 1, \ldots, n;\ j = 1, \ldots, J \right\}. \tag{1}$$

In Figure 1, a bi-dimensional artificial dataset is represented. In this dataset two groups of 50 subjects classified as normotensive (black color with $\mu' = [75, 140]$) and hypertensive (red color with $\mu' = [100, 210]$) have been simulated. In the left plot of Figure 1, the dataset is represented in ordinary form (i.e., with a radius equal to zero), while in the right one, the dataset is represented in interval-valued form (i.e., with a radius bigger than zero). In the literature on data analysis, a great deal of attention is paid to statistical methods to treat interval-valued data, in different research areas [7–9,11–13].

In a classical cluster analysis framework different interesting methods have been suggested. In particular, Ref. [14] proposed a clustering method for symbolic data; Ref. [15] proposed a similarity measure for comparing interval-valued data and a modified agglomerative method for clustering symbolic data. Ref. [16] proposed a partitional dynamic clustering method for interval data based on adaptive Hausdorff distances; Ref. [17] suggested clustering methods for interval data based on single adaptive distances.

However, an interesting line of research has focused on clustering of interval-valued data based on fuzzy approaches, where the weighting exponent m controls the extent of membership sharing between fuzzy clusters [7,18–21]. Ref. [22] remarked that this "strange" parameter is unnatural and has no physical meaning. Then, in the above objective function, we may remove m, but in this case, the procedure cannot generate the membership update Equations [23]. To this purpose, Refs. [22,24] suggested a new approach to fuzzy clustering by proposing the so-called Maximum Entropy Inference Method. The idea underlies the paper by [25] where the trade-off between fuzziness and compactness is dealt with by

introducing a unique objective function reformulating the maximum entropy method in terms of regularization of the fuzzy c-means (FCM) function.

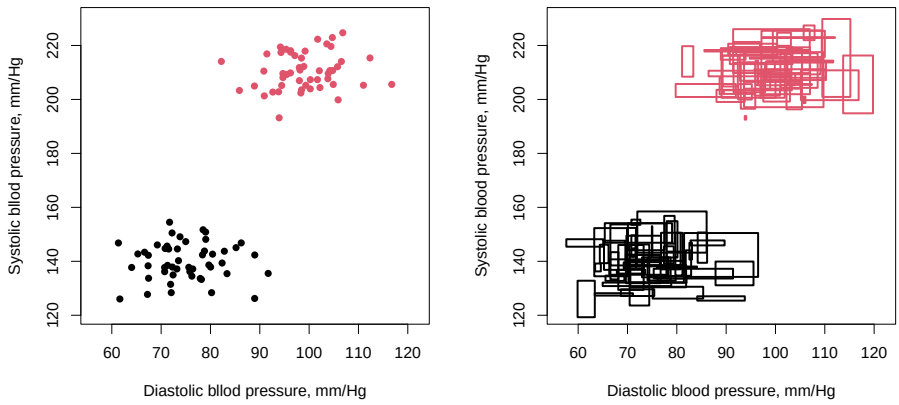


Figure 1. Artificial data generated by two bi-variate Normal distributions. To the left we have a dataset in an ordinary form; to the right, we have an interval-valued dataset.

In the literature, many authors proposed the entropy-based approach as regularization in fuzzy clustering modeling. In particular, Ref. [26] proposed an entropy-based fuzzy clustering method that automatically identifies the number and initial locations of cluster centers. Successively, it removes all data points having a similarity larger than a threshold with the chosen cluster center. The procedure is repeated till all data points are removed; Refs. [27,28] suggested a generalized objective function with additional variables. These authors consider a covariance matrix and show an equivalence between their Kullback–Leibler (KL) fuzzy clustering and the Gaussian mixture model. The method of fuzzy clustering using the KL information is called the entropy-based method of FCM; Ref. [29] suggested an axiomatic derivation of the Maximum Entropy Inference (and also of the possibilistic) clustering approach, based on a unifying principle of physics, that of Extreme Physical Information (EPI) defined by [30]; Ref. [23] suggest fuzzy unsupervised clustering models based on Shannon entropy regularization in order to classify time-varying data; Ref. [31] proposed a new fuzzy clustering method based on FCM and the relative entropy is added to its objective function as a regularization function to maximize the dissimilarity between clusters; Ref. [32] presented an entropy-based FCM segmentation method that incorporates the uncertainty of classification of individual pixels within the classical framework of FCM; Ref. [33] showed a novel method considering noise intelligently based on the existing FCM approach, called adaptive-FCM and its extended version (adaptive-REFCM) in combination with relative entropy; more recently, Ref. [34] proposed an entropy-based regularization approach to fuzzify the partition and to weight features, enabling the method to capture more complex patterns, identify significant features, and yield better performance facing high-dimensional data. Notice that the here-cited proposals on the models with entropy-based regularization, regard applications on ordinary point data.

Following this research line, in this work, an entropy-based fuzzy clustering technique for interval-valued data for cancer detection is suggested. The novelty of this statistical approach is to consider the uncertainty of the data in the classification procedure using the standard deviation of data variables as a measure of the uncertainty. Moreover, the presence of an entropy-based regularization redresses the uncertainty among the statistical units, especially in the boundary region guarantying a more precise classification with respect to the other competitor models. The model is named entropy-based fuzzy clustering for interval-valued data (EFC-ID). Since for all kinds of cancer, it is particularly important to improve the accuracy of early diagnosis, and that conventional early diagnosis mainly relies on human experience, an automatic classification procedure can be improved the cancer detection in screening stages.

The paper is organized as follows: in Section 2 the principal ingredients of EFC-ID

approach are provided; in Section 3 the mathematical structure and the algorithm of the model are described; in Section 4 a detailed simulation study and comparison with other fuzzy and not fuzzy clustering models for interval-valued data is proposed; in Section 5 the results obtained by the EFC-ID application on empirical data are shown; finally, in Section 6 some concluding remarks and the lines for future research in this field are provided.

2. Principal Ingredients

In this section, the principal ingredients of the entropy-based fuzzy clustering approach for interval-valued data (EFC-ID) are provided. The fundamental ingredients of this classification model are *(i)* the dissimilarity/distance measure to consider, and *(ii)* the entropy regularization approach applied in the fuzzy clustering framework.

2.1. Euclidean Distance

The generic interval-valued data pertaining to the i-th observation with respect to the j-th interval-valued feature can be shown as the pair (c_{ij},r_{ij}), $i = 1,\ldots,n$ and $j = 1,\ldots,J$, where c_{ij} denotes the center and r_{ij} the radius of the interval.

In the literature, several metrics have been suggested for interval-valued. In this paper, we adopt a weighted distance measure, proposed by [35]. In this case, the distance between each pair of observations is measured by separately considering the distances for the centers and the radii of the interval-valued data and using a suitable weighting system for such distance components. Formally, by considering the i-th and i'-th observations, we have

$$d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i'}) = \left[w_c^2 d^2(\mathbf{c}_i, \mathbf{c}_{i'}) + w_r^2 d^2(\mathbf{r}_i, \mathbf{r}_{i'})\right]^{\frac{1}{2}}, \tag{2}$$

where $d^2(\mathbf{c}_i, \mathbf{c}_{i'}) = \|\mathbf{c}_i - \mathbf{c}_{i'}\|^2$ is the squared Euclidean distance between the centers and $d^2(\mathbf{r}_i, \mathbf{r}_{i'}) = \|\mathbf{r}_i - \mathbf{r}_{i'}\|^2$ is the squared Euclidean distance between the radii, while w_c and w_r are suitable weight for the center component and the radius component, respectively.

Moreover, we assume the following conditions: (*i*) $w_c + w_r = 1$ (*normalization condition*) and (*ii*) $w_c \geq w_r \geq 0$ (*coherence condition*). In particular, by means of the *coherence condition* we manage to exclude the anomalous case where the radius component, which represents the uncertainty around the centers of the interval-valued data, has more importance than the center component, which represents the core information of each interval-valued datum. Furthermore, through the *normalization condition* we can easily assess, in a comparative fashion, the contributions of the center and radius components in the distance computation.

The distance measure shown in Equation (2) has the following properties:

1. $d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i'})$ is a metric, i.e., the properties of identity, non-negativity, symmetry, and the triangular inequality are satisfied (for details see [8]).
2. $d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i'})$ is computationally easy and theoretically intuitive.
3. $d(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i'})$ tunes suitable the contribution of the (squared) distance measures of the center and radius components of the interval-valued data by means of a weighting system capable to assign objectively (by means of an optimization process) or subjectively (by means of the expertise and experience of the researcher) weights to the two distance components.

2.2. Shannon Entropy Regularization

In this paper, we focus on the entropy regularization approach in a fuzzy clustering framework because is known that the maximum entropy principle, as applied to fuzzy clustering, provides a new perspective to face the problem of fuzzifying the clusterization of the units, while ensuring the maximum of compactness of the obtained clusters [23,33]. The former objective is achieved by maximizing the entropy (i.e., the uncertainty) of the classification of the units into the various clusters. The latter objective is obtained by constraining the above maximization process in such a way as to minimize the overall distance of the units from the cluster prototypes (i.e., to maximize cluster compactness). In other words, we use an entropy-based FCM segmentation method that incorporates the

uncertainty of classification of individuals within the classical framework of FCM [32].

Through this technique, the Shannon entropy measure is employed in the objective function of FCM to redress the uncertainty among the statistical units, especially in the boundary region. Additionally, given the nature of our data (i.e., interval-valued), a weighted distance measure proposed by [35] is adopted. In this case, the distance between each pair of observations is measured by separately considering the distances for the centers and the radii of the interval-valued data and using a suitable weighting system for such distance components.

3. Model and Algorithm

The proposed model has been processed through the statistical software R Studio release 2022.02.0. The algorithm and dataset used in the simulation study and empirical applications are uploaded on the following web page: https://github.com/mfordellone/EFC-ID (accessed on 1 February 2023).

3.1. Optimization Problem

Let $\mathbf{X}$ be a $n \times J$ interval-valued data matrix. Given the distance measure shown in Equation (2), in which we assume that the weights (i.e., w_c and w_r) are objectively computed during the clustering process, we can classify observations within a fuzzy framework, by means of the entropy-based fuzzy clustering (EFC-ID) model, characterized as follows:

$$\begin{aligned} \min \quad & J_{EFC-ID}(\mathbf{U}, \tilde{\mathbf{X}}, w) = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}\left[w_c^2 d^2(\mathbf{c}_i, \mathbf{c}_g) + w_r^2 d^2(\mathbf{r}_i, \mathbf{r}_g)\right] + \\ & + p\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig} log(u_{ig}) \\ \text{s.t.} \quad & \sum_{g=1}^{k} u_{ig} = 1, u_{ig} \geq 0, \\ & w_c \geq w_r \geq 0, w_c + w_r = 1. \end{aligned} \tag{3}$$

where u_{ig} indicates the membership degree of the i-th unit in the g-th cluster; $d^2(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_g)$ is the squared version of Equation (2) between the i-th unit and the centroid in the g-th cluster; $\mathbf{c}_i$ and $\mathbf{r}_i$ are the centers and radii of the i-th unit, respectively; $\mathbf{c}_g$ and $\mathbf{r}_g$ are the centroids of the centers and radii in the g-th cluster, respectively; $p\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig} log(u_{ig})$ is the *fuzzy entropy function*; p is a weight factor, called *degree of fuzzy entropy*, similar to the weight exponent m used in the fuzzy k-means approach and represents the uncertainty associated with each statistical unit which is defined as the Shannon entropy [22–24]. To simplify things, we can set $w_c = (1 - w)$ and $w_r = w$. In this way, the *normalization condition* is satisfied and the *coherence condition* turns into $0 \leq w \leq 0.5$. Then, the objective function became:

$$\begin{aligned} \min \quad & J_{EFC-ID}(\mathbf{U}, \tilde{\mathbf{X}}, w) = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}\left[(1-w)^2 d^2(\mathbf{c}_i, \mathbf{c}_g) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_g)\right] + \\ & + p\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig} log(u_{ig}) \\ \text{s.t.} \quad & \sum_{g=1}^{k} u_{ig} = 1, u_{ig} \geq 0, \\ & 0 \leq w \leq 0.5 \end{aligned} \tag{4}$$

By solving the constrained quadratic minimization problem shown in Equation (4) *via* Lagrangian multiplier method, we obtain the optimal solutions u_{ig} and w. In particular, by considering the following Lagrangian function:

$$\begin{aligned} L_m(u_{ig}, \lambda, w) = \sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}\Big[(1-w)^2 d^2(\mathbf{c}_i, \mathbf{c}_g) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_g)\Big] + \\ + p\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig} log(u_{ig}) - \lambda\left(\sum_{g=1}^{k} u_{ig} - 1\right), \end{aligned} \tag{5}$$

and setting the first partial derivatives with respect u_{ig} and λ equal zero, we obtain

$$\begin{aligned} \frac{\partial L_m(u_{ig}, \lambda, w)}{\partial u_{ig}} = 0 \Leftrightarrow \Big[(1-w)^2 d^2(\mathbf{c}_i, \mathbf{c}_g) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_g)\Big] + \\ + p(log(u_{ig}) + 1) - \lambda = 0, \end{aligned} \tag{6}$$

$$\frac{\partial L_m(u_{ig}, \lambda, w)}{\partial \lambda} = 0 \Leftrightarrow \sum_{g=1}^{k} u_{ig} - 1 = 0. \tag{7}$$

From Equation (6), we obtain

$$log(u_{ig}) = \frac{1}{p}\Big[\lambda - \Big[(1-w)^2 d^2(\mathbf{c}_i, \mathbf{c}_g) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_g)\Big]\Big] - 1, \tag{8}$$

and then

$$u_{ig} = \exp\left[\frac{\lambda}{p} - \frac{1}{p}\Big[(1-w)^2 d^2(\mathbf{c}_i, \mathbf{c}_g) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_g)\Big] - 1\right]. \tag{9}$$

By considering Equation (7):

$$\exp\left(\frac{\lambda}{p} - 1\right) = \frac{1}{\sum_{g=1}^{k}\left[\frac{1}{\exp\left[(1/p)[(1-w)^2 d^2(\mathbf{c}_i, \mathbf{c}_g) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_g)]\right]}\right]}, \tag{10}$$

and by replacing Equation (10) in Equation (9), we obtain

$$u_{ig} = \frac{1}{\sum_{g'=1}^{k}\left[\frac{\exp\left[(1/p)[(1-w)^2 d^2(\mathbf{c}_i, \mathbf{c}_g) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_g)]\right]}{\exp\Big[(1/p)[(1-w)^2 d^2(\mathbf{c}_i, \mathbf{c}_{g'}) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_{g'})]\Big]}\right]}. \tag{11}$$

The normalization condition for w is implicitly satisfied. To take into account the *coherence condition*, observing that Equation (5) is a parabola with respect to w, the optimum value of w results in the minimum between the abscissa of its vertex and 0.5 [8], i.e.,

$$w = \min\left\{\frac{\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}\left[d^2(\mathbf{c}_i, \mathbf{c}_g)\right]}{\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}\left[d^2(\mathbf{c}_i, \mathbf{c}_g) + d^2(\mathbf{r}_i, \mathbf{r}_g)\right]}, 0.5\right\}. \tag{12}$$

Finally, we compute the centroids for the centers and radii through the steps shown in Equations (13) and (14), respectively.

$$\begin{aligned}
&\frac{\partial J_{EFC-ID}(\mathbf{U}, \tilde{\mathbf{X}}, w)}{\partial \mathbf{c}_g} = 0 \Leftrightarrow \\
&\sum_{i=1}^{n} u_{ig} \Big[(1-w)^2 d^2(\mathbf{c}_i, \mathbf{c}_g) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_g) \Big] + p \sum_{i=1}^{n} u_{ig} log(u_{ig}) = 0 \Leftrightarrow \\
&\sum_{i=1}^{n} u_{ig} \Big[(1-w)^2 (\mathbf{c}_i^2 + 2\mathbf{c}_i\mathbf{c}_g + \mathbf{c}_g^2) + w^2 (\mathbf{r}_i^2 + 2\mathbf{r}_i\mathbf{r}_g + \mathbf{r}_g^2) \Big] \\
&+ p \sum_{i=1}^{n} u_{ig} log(u_{ig}) = 0 \Leftrightarrow \\
&\sum_{i=1}^{n} u_{ig} \Big[(1-w)^2 (\mathbf{c}_i + \mathbf{c}_g) \Big] = 0 \Leftrightarrow \\
&\mathbf{c}_g = \frac{\sum_{i=1}^{n} u_{ig} \mathbf{c}_i}{\sum_{i=1}^{n} u_{ig}}.
\end{aligned} \tag{13}$$

$$\begin{aligned}
&\frac{\partial J_{EFC-ID}(\mathbf{U}, \tilde{\mathbf{X}}, w)}{\partial \mathbf{r}_g} = 0 \Leftrightarrow \\
&\sum_{i=1}^{n} u_{ig} \Big[(1-w)^2 d^2(\mathbf{c}_i, \mathbf{c}_g) + w^2 d^2(\mathbf{r}_i, \mathbf{r}_g) \Big] + p \sum_{i=1}^{n} u_{ig} log(u_{ig}) = 0 \Leftrightarrow \\
&\sum_{i=1}^{n} u_{ig} \Big[(1-w)^2 (\mathbf{c}_i^2 + 2\mathbf{c}_i\mathbf{c}_g + \mathbf{c}_g^2) + w^2 (\mathbf{r}_i^2 + 2\mathbf{r}_i\mathbf{r}_g + \mathbf{r}_g^2) \Big] \\
&+ p \sum_{i=1}^{n} u_{ig} log(u_{ig}) = 0 \Leftrightarrow \\
&\sum_{i=1}^{n} u_{ig} \Big[w^2 (\mathbf{r}_i + \mathbf{r}_g) \Big] = 0 \Leftrightarrow \\
&\mathbf{r}_g = \frac{\sum_{i=1}^{n} u_{ig} \mathbf{r}_i}{\sum_{i=1}^{n} u_{ig}}.
\end{aligned} \tag{14}$$

In order to show an example of application we consider the bi-dimensional interval-valued dataset described in Introduction. In Table 1 are shown the mean and variance of centers and radii used to generate 300 observations of a bi-dimensional interval-valued data with a structure of three groups (i.e., 100 observations for each cluster).

Table 1. Clusters mean and variance of an artificial interval-valued dataset.

	Centers				***Radii***		
	Cluster 1	**Cluster 2**	**Cluster 3**		**Cluster 1**	**Cluster 2**	**Cluster 3**
μ_1	0	−10	10	μ_1	0	−3	3
μ_2	−10	10	10	μ_2	−3	3	3
σ_1^2	5	5	5	σ_1^2	2	2	2
σ_2^2	5	5	5	σ_2^2	2	2	2

By applying the EFC-ID model on this dataset, we have the results shown in Figure 2. Then, the centroids of centers and radii have been correctly identified with the Adjusted Rand index (ARI) value equal to 1.

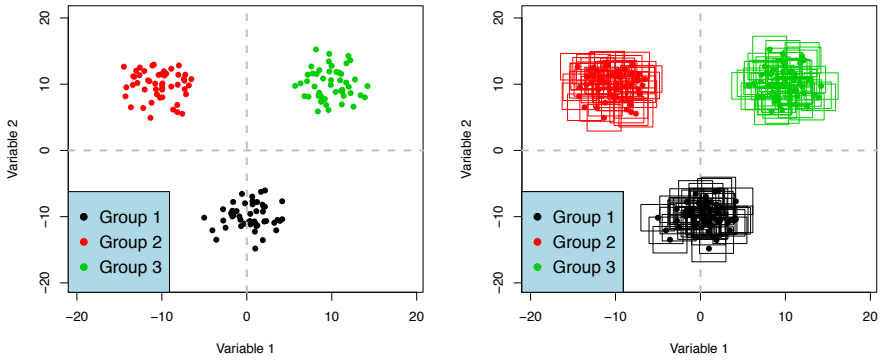


Figure 2. Partition identified by the EFC-ID model of an artificial interval-valued data. The clusters are highlighted through different colors.

3.2. Entropy-Based Fuzzy Clustering Algorithm

In the following, we show the algorithm for the EFC-ID model. Fixed p (*degree of fuzzy entropy*), k (the number of clusters) and *maxiter* (the maximum number of iterations), and set iter = 0, the EFC-ID Algorithm 1 is composed of the following steps:

Algorithm 1 Entropy-based fuzzy clustering algorithm.

1: Randomly generate the membership matrix $\mathbf{U}^{iter}$ subject to constraints shown in (4); **iter = iter+1**
2: Given $\mathbf{U}^{iter-1}$, compute the centroids for the centers and radii $\mathbf{C}^{iter-1}$ and $\mathbf{R}^{iter-1}$;
3: Compute w^{iter-1} according Equation (12);
4: Update the membership matrix $\mathbf{U}^{iter}$ according Equation (11);
5: **if** $\|\mathbf{U}^{iter} - \mathbf{U}^{iter-1}\| > \epsilon$ & **iter** $<$ maxiter ; go to step 2.
6: **else** exit loop.
7: **Return**: the membership matrix $\mathbf{U}$,
the centroids for the centers and radii $\mathbf{C}$ and $\mathbf{R}$,
the weight w,
the number of iteration *iter*.

Notice that, given the constraints on $\mathbf{U}$, the algorithm can be expected to be rather sensitive to *local optima*. For this reason, it is recommended the use of some randomly started runs to find the best solution.

3.3. Cluster Validity Indices

The first step of the EFC-ID application is the choice of the optimal *degree of fuzzy entropy*. For this purpose, two cluster validity indices are considered: the *partition coefficient* index (V_{PC}) and the *partition entropy* measure (V_{PE}). The former one can be viewed as a mean over the n units of Onicescu's *information energy* [36] in a fuzzy setting:

$$V_{PC} = \frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig}^2, \tag{15}$$

the latter one is the same measure in an entropy-based setting [23]:

$$V_{PE} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{k} u_{ig} log(u_{ig}). \tag{16}$$

Both V_{PC} and V_{PE} measure the degree of the overlapping among clusters. Moreover, V_{PC} is a decreasing function of p in the fuzzy entropy objective function, while V_{PE} is an increasing function of this parameter.

Then, given the number of clusters k, an optimal value of p is the value at which $V_{PC} = V_{PE}$, obtaining a good compromise between maximizing the separation of clusters (i.e., the V_{PC} minimization), and optimizing the *fuzziness* degree of classification (i.e., the V_{PE} maximization). This criterion can be used to choose the optimal number of clusters also.

4. Simulation Study

To investigate the performance of the entropy-based fuzzy clustering (EFC-ID) model, a simulation study has been carried out. The aim of this simulation study is to study the behavior of EFC-ID in different cases that could be occurred in empirical applications (e.g., well-separated and not well-separated clusters, presence of fuzzy points, groups structure applied on centers, on radii, or on both, etc.). In particular, is very interesting to study the EFC-ID model in terms of weights (w) estimate and identification of the natural partition for different *degrees of entropy*.

The proposed model has also been compared with other fuzzy clustering models for interval-valued data, i.e., fuzzy k-means clustering (FKM) proposed by [35] and fuzzy relational clustering (FRC), and with other crisp models, i.e., hierarchical clustering (HC). Moreover, EFC-ID has been also compared with another version of the entropy-based clustering model, i.e., the entropy-based fuzzy clustering model for point data (here called EFC) and the entropy-based clustering for interval-valued data (EC-ID).
For FRC and HC a dissimilarity measure for interval-valued data based on OR proposed by [37] has been used.

Regarding the simulation scheme, three data generation scenarios have been considered. In each scenario, the simulated dataset is constructed in such a way that two well-separated clusters ($k = 2$) with the same size are generated (i.e., cluster 1: $1, \ldots, n/2$, cluster 2: $n/2 + 1, \ldots, n$). Following the simulation line proposed by [8,19,21], we have:

- *centers-radii scenario*, where the centers and the radii of the interval-valued data generated have a group structure.
- *centers scenario*, where the radii of the interval-valued data are all randomly generated, while the centers of the data generated have a group structure.
- *radii scenario*, where the centers of the interval-valued data are all randomly generated, while the radii of the data generated have a group structure.

In Table 2, the details on the simulated scheme are shown.

Table 2. Data generation in the simulation scheme.

Scenario	Centers	Radii
centers-radii	Cluster 1: U[0, 1]	Cluster 1: U[0, 1]
	Cluster 2: U[3.5, 4.5]	Cluster 2: U[1.5, 2.5]
centers	Cluster 1: U[0, 1]	Cluster 1: U[0, 2]
	Cluster 2: U[3.5, 4.5]	Cluster 2: U[0, 2]
radii	Cluster 1: U[0, 2]	Cluster 1: U[0, 1]
	Cluster 2: U[0, 2]	Cluster 2: U[1.5, 2.5]

Each simulated dataset is composed of one hundred objects ($n = 100$) and two interval-valued variables ($J = 2$). Moreover, for the purpose of evaluating the fuzzy clustering performances of the proposed model in presence of fuzzy points (i.e., data points with memberships degree for each cluster equal to 0.5), three different percentages of fuzzy points (1, 5, and 10%) have been included in the 100 objects. In this way, we have 4 different datasets for each scenario. Note that for each scenario the data-generating process has been replicated 300 times. Finally, we have also set three values of the fuzziness parameter m (1.1, 1.5, and 2, respectively) and three values of the fuzzy entropy parameter p (0.10, 0.20,

and 0.40), to detect how the clustering performance is affected by these parameters. For the hierarchical clustering model, HC (i.e., hard clustering), the *single linkage* and the *complete linkage* approaches have been considered.

For evaluating the performance of the model the Frobenius distance (Fdist) computed between the natural (generated) memberships matrix $\mathbf{U}_c$ and the memberships matrix $\hat{\mathbf{U}}$ obtained by the model, has been used. This approach is often used as a stopping rule in some fuzzy clustering algorithms [38]. The Frobenius distance has been then averaged over the 300 simulation runs.

The results are presented in Table 3 with respect to different percentages of fuzzy points, different fuzziness/fuzzy entropy parameters, and different linkage methods. Table 3 shows that the average values of Fdist recorded for EFC-ID and HC are exactly equal to zero in the case of well-separated clusters (i.e., the natural partition is correctly identified by the model), whereas FKM shows Fdist values slightly higher. Another remarkable finding is that the clustering performance of our proposed model is slightly affected by the percentage of fuzzy points with respect to the other models, especially when the *degree of fuzzy entropy* increases. Moreover, EFC-ID shows better performance than all the other approaches especially in the *radii scenario*. Notice that all the results of the not fuzzy approaches can be compared with EFC-ID when $p = 0.2$ (i.e., the medium *fuzziness degree*).

Table 3. Clustering models performance with well-separated clusters (0% of fuzzy points) and not well-separated clusters (1%, 5%, 10% of fuzzy points).

		Centers-Radii		*Centers*		*Radii*	
	Fuzzy Points	**Fdist**	**w**	**Fdist**	**w**	**Fdist**	**w**
Entropy-based fuzzy clustering for interval-valued data (EFC-ID)							
$p = 0.10$	0%	0.000	0.490	0.000	0.200	0.023	0.500
	1%	0.000	0.500	0.158	0.254	0.140	0.500
	5%	1.578	0.500	1.581	0.380	0.665	0.500
	10%	2.236	0.500	2.236	0.445	1.664	0.500
$p = 0.20$	0%	0.000	0.490	0.000	0.200	0.522	0.492
	1%	0.014	0.500	0.043	0.254	0.513	0.493
	5%	1.480	0.500	1.544	0.381	0.580	0.496
	10%	2.231	0.500	2.232	0.445	0.902	0.500
$p = 0.40$	0%	0.000	0.490	0.000	0.200	7.062	0.338
	1%	0.006	0.500	0.017	0.254	7.028	0.340
	5%	0.996	0.500	1.205	0.388	6.887	0.350
	10%	2.092	0.500	2.097	0.452	6.707	0.364
Fuzzy *k*-means clustering (FKM)							
$m = 1.10$	0%	0.000	0.490	0.000	0.200	0.021	0.500
	1%	0.016	0.500	0.171	0.254	0.284	0.500
	5%	1.593	0.500	1.623	0.382	1.275	0.500
	10%	2.236	0.500	2.239	0.445	2.178	0.500
$m = 1.50$	0%	0.002	0.492	0.005	0.202	0.632	0.500
	1%	0.014	0.500	0.048	0.257	0.641	0.500
	5%	1.556	0.500	1.729	0.407	0.730	0.500
	10%	2.475	0.500	2.527	0.490	1.112	0.500
$m = 2.00$	0%	0.176	0.500	0.533	0.500	7.071	0.338
	1%	0.178	0.500	0.540	0.500	7.036	0.340
	5%	0.995	0.500	1.603	0.500	6.892	0.350
	10%	2.245	0.500	2.574	0.500	6.708	0.364

Table 3. *Cont.*

		Centers-Radii		*Centers*		*Radii*	
	Fuzzy Points	**Fdist**	**w**	**Fdist**	**w**	**Fdist**	**w**
Fuzzy relational clustering (FRC)							
$m = 1.10$	0%	5.584	-	5.670	-	5.257	-
	1%	5.815	-	5.630	-	5.315	-
	5%	5.773	-	5.606	-	5.275	-
	10%	5.769	-	5.480	-	5.955	-
$m = 1.50$	0%	0.125	-	3.428	-	4.433	-
	1%	0.496	-	2.565	-	4.232	-
	5%	1.554	-	2.765	-	4.112	-
	10%	2.428	-	2.548	-	4.233	-
$m = 2.00$	0%	0.796	-	1.673	-	8.342	-
	1%	0.827	-	1.842	-	8.661	-
	5%	1.234	-	1.662	-	7.844	-
	10%	2.139	-	1.992	-	7.645	-
Hierarchical clustering (HC)							
Single linkage	0%	0.000	-	1.400	-	2.144	-
	1%	0.707	-	3.659	-	2.558	-
	5%	1.581	-	8.483	-	2.799	-
	10%	2.283	-	9.105	-	2.721	-
Complete linkage	0%	0.000	-	0.009	-	2.144	-
	1%	0.707	-	0.723	-	2.558	-
	5%	1.581	-	1.613	-	2.799	-
	10%	2.236	-	2.286	-	2.721	-

Concerning the weights, we can note that $wc = ws = 0.5$ in the *centers-radii scenario* (i.e., when the variability of data is balanced between centers and radii) and in the *radii scenario* (i.e., when the variability of radii is higher than the variability of centers and then, ws assume the maximum value). Finally, $wc \geq ws$ when the variability of centers is higher than the variability of radii (i.e., in the *centers scenario*).

In order to complete the evaluation of the proposed model, we have compared the EFC-ID results with the entropy-based fuzzy clustering model for point data (EFC) and the entropy-based clustering (EC-ID) model (i.e., the crisp version for interval-valued data). In Table 4, all the results are reported. The results in Table 4 show that the EFC-ID performance is better also than other entropy-based clustering models.

Table 4. Clustering models performance with well-separated clusters (0% of fuzzy points) and not well-separated clusters (1%, 5%, 10% of fuzzy points).

		Centers-Radii		*Centers*		*Radii*	
	Fuzzy Points	**Fdist**	**w**	**Fdist**	**w**	**Fdist**	**w**
Entropy-based fuzzy clustering for interval-valued data (EFC-ID)							
$p = 0.10$	0%	0.000	0.490	0.000	0.200	0.023	0.500
	1%	0.000	0.500	0.158	0.254	0.140	0.500
	5%	1.578	0.500	1.581	0.380	0.665	0.500
	10%	2.236	0.500	2.236	0.445	1.664	0.500
$p = 0.20$	0%	0.000	0.490	0.000	0.200	0.522	0.492
	1%	0.014	0.500	0.043	0.254	0.513	0.493
	5%	1.480	0.500	1.544	0.381	0.580	0.496
	10%	2.231	0.500	2.232	0.445	0.902	0.500
$p = 0.40$	0%	0.000	0.490	0.000	0.200	7.062	0.338
	1%	0.006	0.500	0.017	0.254	7.028	0.340
	5%	0.996	0.500	1.205	0.388	6.887	0.350
	10%	2.092	0.500	2.097	0.452	6.707	0.364

Table 4. *Cont.*

		Centers-Radii		*Centers*		*Radii*	
	Fuzzy Points	**Fdist**	**w**	**Fdist**	**w**	**Fdist**	**w**
Entropy-based fuzzy clustering for point data (EFC)							
$p = 0.10$	0%	0.000	-	0.000	-	9.387	-
	1%	0.547	-	0.547	-	9.360	-
	5%	1.581	-	1.581	-	9.269	-
	10%	2.236	-	2.236	-	9.145	-
$p = 0.20$	0%	0.000	-	0.000	-	9.136	-
	1%	0.100	-	0.100	-	9.108	-
	5%	1.581	-	1.581	-	9.015	-
	10%	2.236	-	2.236	-	8.886	-
$p = 0.40$	0%	0.000	-	0.000	-	8.480	-
	1%	0.030	-	0.030	-	8.452	-
	5%	1.572	-	1.572	-	8.348	-
	10%	2.236	-	2.236	-	8.213	-
Entropy-based clustering for interval-valued data (EC-ID)							
	0%	0.000	0.490	0.000	0.200	0.566	0.500
	1%	0.706	0.500	0.707	0.254	0.625	0.500
	5%	1.581	0.500	1.581	0.380	1.571	0.500
	10%	2.236	0.500	2.236	0.445	2.236	0.500

5. Empirical Applications

In this section, we show the results obtained by the entropy-based fuzzy clustering (EFC-ID) model in two empirical applications. For replication purposes, the reader can refer to the R-scripts and datasets uploaded on the following web page: https://github.com/mfordellone/EFC-ID (accessed on 1 February 2023). Notice that for the classification we assume that the observed diagnosis groups are unknown (i.e., unsupervised classification), and subsequently, we use the natural partitions of datasets to evaluate the diagnostic performance of the models.

5.1. Breast Cancer Wisconsin Data

In this subsection an analysis of the Breast Cancer Wisconsin (Diagnostic) dataset is performed (https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic), accessed on 1 February 2023). This data set was created by [39] and it has been very used for training of statistical methods (e.g., [40]). The endpoint of this statistical analysis is to unsupervised classify the kind of breast cancer (i.e., benign or malignant). The dataset consists of 569 patients: 357 with benign diagnosis and 212 with malignant status. Table 5 shows the features average by a breast cancer diagnosis.

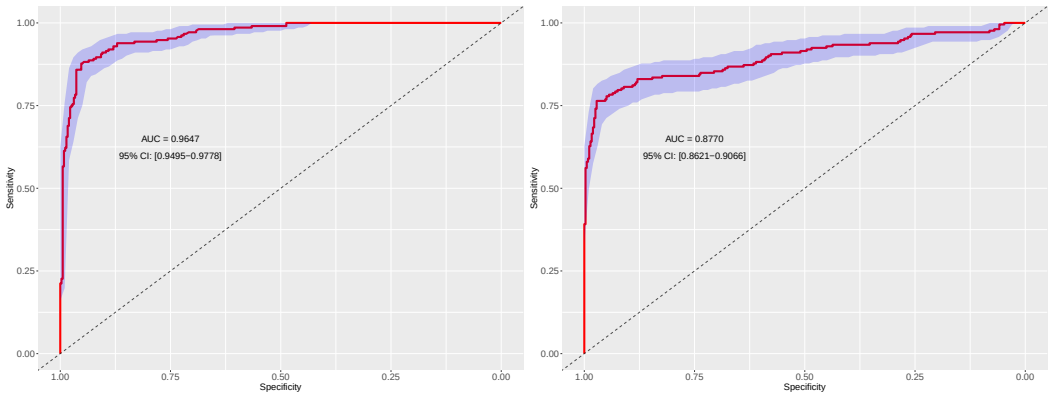
Table 5. Breast Cancer Wiscosin: features average by diagnosis.

	Benign (n = 357)	**Malignant (n = 212)**	***p*-Value**
radius	12.1 ± 1.78	17.5 ± 3.20	<0.001
texture	17.9 ± 4.00	21.6 ± 3.78	<0.001
perimeter	78.1 ± 11.8	115 ± 22	<0.001
area	463 ± 134	978 ± 368	<0.001
smoothness	0.09 ± 0.01	0.10 ± 0.01	<0.001
compactness	0.08 ± 0.03	0.15 ± 0.05	<0.001
concavity	0.05 ± 0.04	0.16 ± 0.07	<0.001
concave points	0.02 ± 0.01	0.09 ± 0.03	<0.001
symmetry	0.17 ± 0.02	0.19 ± 0.03	<0.001
fractal dimension	0.06 ± 0.01	0.06 ± 0.01	0.880

For each feature we have the average ± standard deviation.

In order to include the variability/uncertainty that characterizes the dataset in the classification procedure, the radii have been fixed equal to the standard deviation of data-

features and then the EFC-ID model for malignant breast cancer diagnosis has been applied to the obtained interval-valued dataset. For comparison purposes, also the conventional fuzzy k-means clustering (FKM) for interval-valued data has been applied. This is because FKM is the major competitor model of EFC-ID, showing the best results in the simulation study. Figure 3a,b show the ROC curves obtained by the EFC-ID and FKM approaches, respectively. From Figure 3 we can see that EFC-ID outperforms the FKM with the AUC value equal to 0.9647 (CI 95% 0.9495–0.9778) against the 0.8770 (CI 95% 0.8621–0.9066) obtained by FKM.



(**a**) (**b**)

Figure 3. ROC curves obtained by the EFC-ID and FKM approaches. (**a**) ROC curve obtained by EFC-ID approach; (**b**) ROC curve obtained by FKM approach.

Table 6 shows the performance indexes computed on the results obtained by the two approaches. Notice that, in this application, the optimal entropy parameter p has been fixed to 0.5.

Table 6. Performance indexes obtained by EFC-ID and FKM for malignant breast cancer diagnosis.

	EFC-ID			**FKM**		
	Estimate	Lower *	Upper *	Estimate	Lower *	Upper *
Sensitivity	0.896	0.847	0.934	0.643	0.596	0.738
Specificity	0.934	0.895	0.955	0.916	0.903	0.977
Pos.Pred.Val.	0.864	0.811	0.906	0.898	0.821	0.963
Neg.Pred.Val.	0.937	0.906	0.960	0.666	0.614	0.743
Youden index	0.812	0.762	0.862	0.546	0.411	0.658
Accuracy	0.909	0.882	0.931	0.845	0.800	0.877
Error rate	0.091	0.069	0.118	0.177	0.132	0.188

* 95% exact confidence interval.

From the performance table, we can see that EFC-ID showed better performance than FKM in terms of global performance, Sensibility, Specificity, and Negative Predictive Value. However, the Specificity and the Positive Predictive Value obtained by FKM are better but, unfortunately, also the false negative rate increases. In this case, it seems that FKM unbalances the weights in favor of *centers* ($w_c = 0.97$), providing a result similar to the simple FCM for point data (i.e., when $w_c = 1$), and showing a high rate of false negative events. Conversely, the entropy regularization of EFC-ID guarantees a better balancing of weights ($w_c = 0.65$). Moreover, we think that the high rate of false negative events obtained by FKM is a serious problem in the cancer detection field. In particular, false negative event tests at diagnosis of early disease and of relapse resulted in diagnostic and therapeutic delays.

5.2. Multiple Myeloma Data

In this subsection, an analysis of the Multiple Myeloma Data available in survminer R-package is performed (https://cran.r-project.org/web/packages/survminer/survminer.pdf, accessed on 1 February 2023). Multiple Myeloma data is extracted from publicly available gene expression data (GEO Id: GSE4581). The endpoint of this statistical analysis is to unsupervised classify the molecular groups of Multiple Myeloma. The original dataset consists of 256 patients with the IMWG molecular cytogenetic classification as shown in Table 7. However, since there is a large number of groups, we have selected a sub-sample including in the analysis only three molecular groups: Hyperdiploid, MAF, and MMSET (i.e., 131 patients). In this dataset, we have six features that correspond to six different gene expressions (i.e., CCND1, CRIM1, DEPDC1, IRF4, TP53, and WHSC1). Table 8 shows the gene expression average by molecular groups of Multiple Myeloma.

Table 7. Multiple Myeloma Data: IMWG molecular cytogenetic classification.

Molecular Group	# Patients
Cyclin D-1	22
Cyclin D-2	43
Hyperdiploid	66
Low bone disease	31
MAF	21
MMSET	44
Proliferation	29

Table 8. Multiple Myeloma Data: gene expression average by molecular groups.

	Hyperdiploid (n = 66)	MAF (n = 21)	MMSET (n = 44)	*p*-Value
CCND1	1280 ± 1500	434 ± 765	427 ± 602	<0.001
CRIM1	73 ± 174	59 ± 91	482 ± 392	<0.001
DEPDC1	169 ± 111	382 ± 356	311 ± 236	0.006
IRF4	14,500 ± 3850	12,700 ± 4930	12,300 ± 3950	0.013
TP53	1880 ± 669	1350 ± 1040	1240 ± 475	<0.001
WHSC1	139 ± 136	481 ± 454	8100 ± 4560	<0.001

For each feature we have the average ± standard deviation.

Furthermore, in this case, in order to include the variability/uncertainty that characterizes the dataset in the classification procedure, the radii have been fixed equal to the standard deviation of data-features and then the EFC-ID model has been applied for molecular group identification. Such as the previous subsection, the conventional fuzzy *k*-means (FKM) for interval-valued data has been applied for comparison purposes. Notice that, in this application, the optimal entropy parameter p has been fixed to 0.25. In order to evaluate the performance of both approaches, three different evaluation criteria have been used (i.e., the Chi-Squared Index [41], the Accuracy rate and the Adjusted Rand Index [42]). The results obtained are shown in Table 9.

Table 9. Performance indexes obtained by EFC-ID and FKM for Multiple Myeloma classification.

	EFC-ID	FKM
Chi-Squared	91.636	88.257
Accuracy	0.710	0.601
Adjusted Rand Index	0.329	0.211

Furthermore, in this case, the three evaluation criteria show an EFC-ID performance gain than the conventional FKM. In particular, both the empirical applications show that FKM, unlike EFC-ID, unbalances the weights in favor of centers ($wc = 0.97$), providing a result similar to the simple FCM for point data (i.e., when $wc = 1$), and then neglecting the information included in the uncertainty of data. Therefore, the important contribution

of this proposal is the use of entropy regularization which improves the separability of groups and their homogeneity without neglecting the degree of uncertainty characterized by the data.

6. Concluding Remarks

Following a fuzzy approach, in this paper, a new fuzzy clustering technique for interval-valued data is suggested. In particular, by considering a suitable weighted distance, we propose a fuzzy clustering model with entropy regularization (i.e., the EFC-ID model). Moreover, an approach has been proposed where the uncertainty that characterizes the data has been considered in the classification procedure using the degree of variability to estimate it.

The principal advantages of this approach consist of *(i)* the use of entropy regularization approach in a fuzzy clustering framework is the maximum entropy principle that provides a new perspective to facing the problem of fuzzifying the clusterization of the units while ensuring the maximum of compactness of the obtained clusters; *(ii)* including the uncertainty of the data in the classification procedure leads more homogeneous and separated groups partitions; *(iii)* the multi-group approach facilitates the use of this approach for other purposes as stages detection, response classes identification, prognosis classification, etc.); *(iv)* the external procedure of uncertainty recognition leads to fix a different kind of interval measures (e.g., specific percentile differences); *(v)* the weighted distance guarantees that the point data has a bigger weight than the radii. In this way, the risk to associate the biggest relevance to the uncertainty in the classification procedure is reduced.

Conversely, the principal disadvantages consist of *(i)* the unknown membership function of the imprecise data; *(ii)* the approach could suffer in the case of a small sample size which inflates the radii.

To investigate the performance and effectiveness of the proposed model a simulation study has been carried out. In particular, the aim of the simulation study has been to study the behavior of the EFC-ID model in terms of weights (w) estimate and identification of the natural partition for different degrees of entropy in different cases that could be occurred in empirical applications: *(i)* well-separated clusters, *(ii)* not well-separated clusters (i.e., when fuzzy points there are in the data structure), *(iii)* groups structure applied on centers only, on the radii only, or on both. Results have shown that the proposed approach is more able to distinguish the natural clusters as well as to identify prototypes with respect to other methodologies. The proposed model has been compared with crisp and fuzzy models for interval-valued data. We have also analyzed two real case studies. In all cases, the proposed approach has shown good performance in identifying the natural partition and the advantages of the use of EFC-ID have been detailed illustrated.

For future research could be interesting to embed a cross-validation approach in EFC-ID to select different uncertainty measures with respect to the standard deviation (e.g., inter-quartile range) in order to obtain also non-symmetrical imprecise data (e.g., trapezoidal data [43].

Author Contributions: Conceptualization, M.F.; methodology, M.F., I.D.B. and P.C.; software, M.F. and I.D.B.; validation, D.B. and P.C.; formal analysis, M.F., D.B. and P.C.; writing—original draft preparation, P.C.; writing—review and editing, P.C.; visualization, M.F. and I.D.B.; supervision, P.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fordellone, M.; Chiodini, P. Unsupervised Hierarchical Classification Approach for Imprecise Data in the Breast Cancer Detection. *Entropy* **2022**, *24*, 926. [CrossRef] [PubMed]
2. Oosterhuis, W.P.; Bayat, H.; Armbruster, D.; Coskun, A.; Freeman, K.P.; Kallner, A.; Koch, D.; Mackenzie, F.; Migliarino, G.; Orth, M.; et al. The use of error and uncertainty methods in the medical laboratory. *Clin. Chem. Lab. Med. (CCLM)* **2018**, *56*, 209–219. [CrossRef] [PubMed]
3. Analytical Methods Committee. Uncertainty of measurement: Implications of its use in analytical science. *Analyst* **1995**, *120*, 2303–2308. [CrossRef]
4. White, G.H.; Farrance, I. Uncertainty of measurement in quantitative medical testing: A laboratory implementation guide. *Clin. Biochem. Rev.* **2004**, *25*, S1. [PubMed]
5. Farrance, I.; Frenkel, R. Uncertainty of measurement: A review of the rules for calculating uncertainty components through functional relationships. *Clin. Biochem. Rev.* **2012**, *33*, 49. [PubMed]
6. Strout, T.D.; Hillen, M.; Gutheil, C.; Anderson, E.; Hutchinson, R.; Ward, H.; Kay, H.; Mills, G.J.; Han, P.K. Tolerance of uncertainty: A systematic review of health and healthcare-related outcomes. *Patient Educ. Couns.* **2018**, *101*, 1518–1537. [CrossRef]
7. Denoeux, T.; Masson, M. Multidimensional scaling of interval-valued dissimilarity data. *Pattern Recognit. Lett.* **2000**, *21*, 83–92. [CrossRef]
8. D'Urso, P.; De Giovanni, L. Robust clustering of imprecise data. *Chemom. Intell. Lab. Syst.* **2014**, *136*, 58–80. [CrossRef]
9. D'Urso, P.; Leski, J. Fuzzy c-ordered medoids clustering for interval-valued data. *Pattern Recognit.* **2016**, *58*, 49–67. [CrossRef]
10. D'Urso, P.; De Giovanni, L. Midpoint radius self-organizing maps for interval-valued data with telecommunications application. *Appl. Soft Comput.* **2011**, *11*, 3877–3886. [CrossRef]
11. Coppi, R.; Giordani, P.; D'Urso, P. Component models for fuzzy data. *Psychometrika* **2006**, *71*, 733. [CrossRef]
12. D'Urso, P.; Giordani, P. A possibilistic approach to latent component analysis for symmetric fuzzy data. *Fuzzy Sets Syst.* **2005**, *150*, 285–305. [CrossRef]
13. Giordani, P.; Kiers, H.A. Principal component analysis of symmetric fuzzy data. *Comput. Stat. Data Anal.* **2004**, *45*, 519–548. [CrossRef]
14. Gowda, K.C.; Diday, E. Symbolic clustering using a new dissimilarity measure. *Pattern Recognit.* **1991**, *24*, 567–578. [CrossRef]
15. Guru, D.; Kiranagi, B.B.; Nagabhushan, P. Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns. *Pattern Recognit. Lett.* **2004**, *25*, 1203–1213. [CrossRef]
16. De Carvalho, F.d.A.; Lechevallier, Y. Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern Recognit.* **2009**, *42*, 1223–1236. [CrossRef]
17. De Carvalho, F.d.A.; de Souza, R.M.; Chavent, M.; Lechevallier, Y. Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognit. Lett.* **2006**, *27*, 167–179. [CrossRef]
18. De Carvalho, F.d.A.; Tenório, C.P. Fuzzy K-means clustering algorithms for interval-valued data based on adaptive quadratic distances. *Fuzzy Sets Syst.* **2010**, *161*, 2978–2999. [CrossRef]
19. D'Urso, P.; De Giovanni, L.; Massari, R. Trimmed fuzzy clustering for interval-valued data. *Adv. Data Anal. Classif.* **2015**, *9*, 21–40. [CrossRef]
20. D'Urso, P.; Giordani, P. A robust fuzzy k-means clustering model for interval valued data. *Comput. Stat.* **2006**, *21*, 251–269. [CrossRef]
21. D'Urso, P.; Massari, R.; De Giovanni, L.; Cappelli, C. Exponential distance-based fuzzy clustering for interval-valued data. *Fuzzy Optim. Decis. Mak.* **2017**, *16*, 51–70. [CrossRef]
22. Li, R.P.; Mukaidono, M. A maximum-entropy approach to fuzzy clustering. In Proceedings of the 1995 IEEE International Conference on Fuzzy Systems, Yokohama, Japan, 20–24 March 1995; Volyme 4, pp. 2227–2232.
23. Coppi, R.; D'Urso, P. Fuzzy unsupervised classification of multivariate time trajectories with the Shannon entropy regularization. *Comput. Stat. Data Anal.* **2006**, *50*, 1452–1477. [CrossRef]
24. Li, R.P.; Mukaidono, M. Gaussian clustering method based on maximum-fuzzy-entropy interpretation. *Fuzzy Sets Syst.* **1999**, *102*, 253–258. [CrossRef]
25. Sadaaki, M.; Masao, M. Fuzzy c-means as a regularization and maximum entropy approach. In Proceedings of the 7th International Fuzzy Systems Association World Congress (IFSA'97), Prague, Czech Republic, 25–30 June 1997.
26. Yao, J.; Dash, M.; Tan, S.; Liu, H. Entropy-based fuzzy clustering and fuzzy modeling. *Fuzzy Sets Syst.* **2000**, *113*, 381–388. [CrossRef]
27. Ichihashi, H. Gaussian mixture PDF approximation and fuzzy c-means clustering with entropy regularization. In Proceedings of the 4th Asian Fuzzy Systems Symposium, Tsukuba, Japan, 31 May–3 June 2000; pp. 217–221.
28. Miyagishi, K.; Yasutomi, Y.; Ichihashi, H.; Honda, K. Fuzzy Clustering with regularization by KL information. In Proceedings of the 16th Fuzzy System Symposium, Akita, Japan, 6–8 September 2000; pp. 549–550.
29. Ménard, M.; Eboueya, M. Extreme physical information and objective function in fuzzy clustering. *Fuzzy Sets Syst.* **2002**, *128*, 285–303. [CrossRef]
30. Frieden, B.R. Physics from Fisher information: A unification. *Am. J. Phys.* **2000**, *68*, 1064. [CrossRef]
31. Zarinbal, M.; Zarandi, M.F.; Turksen, I. Relative entropy fuzzy c-means clustering. *Inf. Sci.* **2014**, *260*, 74–97. [CrossRef]

32. Kahali, S.; Sing, J.K.; Saha, P.K. A new entropy-based approach for fuzzy c-means clustering and its application to brain MR image segmentation. *Soft Comput.* **2019**, *23*, 10407–10414. [CrossRef]
33. Gao, Y.; Wang, D.; Pan, J.; Wang, Z.; Chen, B. A novel fuzzy c-means clustering algorithm using adaptive norm. *Int. J. Fuzzy Syst.* **2019**, *21*, 2632–2649. [CrossRef]
34. Ashtari, P.; Haredasht, F.N.; Beigy, H. Supervised fuzzy partitioning. *Pattern Recognit.* **2020**, *97*, 107013. [CrossRef]
35. D'Urso, P.; Giordani, P. A weighted fuzzy c-means clustering model for fuzzy data. *Comput. Stat. Data Anal.* **2006**, *50*, 1496–1523. [CrossRef]
36. Lepădatu, C.; Nitulescu, E. Information energy and information temperature for molecular systems. *Acta Chim. Slov* **2003**, *50*, 539–546.
37. Kabir, S.; Wagner, C.; Havens, T.C.; Anderson, D.T.; Aickelin, U. Novel similarity measure for interval-valued data based on overlapping ratio. In Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy , 9–12 July 2017; pp. 1–6.
38. Leski, J. Towards a robust fuzzy clustering. *Fuzzy Sets Syst.* **2003**, *137*, 215–233. [CrossRef]
39. Mangasarian, O.L.; Street, W.N.; Wolberg, W.H. Breast cancer diagnosis and prognosis via linear programming. *Oper. Res.* **1995**, *43*, 570–577. [CrossRef]
40. Agarap, A.F.M. On breast cancer detection: An application of machine learning algorithms on the wisconsin diagnostic dataset. In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing, Phuoc Island, Vietnam, 2–4 February 2018; pp. 5–9.
41. Jin, X.; Xu, A.; Bie, R.; Guo, P. Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In *International Workshop on Data Mining for Biomedical Applications*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 106–115.
42. Steinley, D. Properties of the Hubert-Arable Adjusted Rand Index. *Psychol. Methods* **2004**, *9*, 386. [CrossRef]
43. Hüllermeier, E. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *Int. J. Approx. Reason.* **2014**, *55*, 1519–1534. [CrossRef]