Assessing clustering methods using Shannon's entropy

Anis Hoayek, Didier Rullière

Assessing clustering methods using Shannon's entropy

Anis Hoayek* and Didier Rullière*

September 15, 2022

Abstract

Unsupervised clustering algorithms are a very important source of information for how a dataset may be classified into subgroups of homogeneous sets. For these algorithms we present a full analysis of their quality by introducing a clustering confidentness metric. Based on this metric, a statistical test and a correction of cluster probabilities are introduced to improve the performance of the different algorithms. These results are illustrated by simulation analysis and by an application on a real world data set.

1 Introduction

In unsupervised learning, clustering methods are very popular methods that associate each observation to a cluster index. Such methods are known as hard clustering approaches. In situations where some observations are difficult to associate with certainty to specific clusters, fuzzy clustering may be used: fuzzy clustering methods, known also as soft clustering methods, associate to each observation the probability to belong to each possible cluster index, (see e.g Ruspini et al., 2019; Yang, 1993; De Oliveira and Pedrycz, 2007, among many other references). There is a wide diversity of clustering methods: a first subdivision is to distribute clustering methods into two families: 1) partitional clustering algorithm (e.g. K-means, density based clustering, genetic algorithm and many other methods), (see e.g. MacQueen (1967); Kriegel et al. (2011); Forrest (1996); among many other references) in which data is organized into a sequence of groups without any hierarchical structure (Ezugwu et al. (2022)); 2) hierarchical clustering algorithm (e.g. Linkage algorithm, divisive clustering), (see e.g. Dawyndt et al. (2005); Roux (2015)) .

A problem for fuzzy clustering is to compare different available methods. Are they overconfident, as in the case where proposed probabilities are all close to 0

*Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F - 42023 Saint-Etienne France

or 1? are they underconfident, as in the case where each observation can belong to any cluster, with equal probabilities? how to measure this confidentness of the clustering method? how to compare clustering methods? how to propose corrections to proposed probabilities in case of over or underconfident clustering method?

The confidentness of a clustering method is indeed of practical interest, beyond the quality of a classification. If the clustering method is too uncertain, it may induce avoidable checks (medical investigations for example, with costs or adverse effects). If the clustering method is too confident, it may disable useful alerts (need of medical checks for example). It may be interesting also to identify points in proposed clusters where the clustering is too uncertain. It is thus of interest to choose carefully the fuzziness of a clustering method.

In the literature, to our knowledge there is very few works on the fuzziness assessment of a clustering method. Some existing indices, as Dave's, Bezdek, and Xi-Beni validity measurement indices (see for example Bataineh et al., 2011), do not directly assess the confidentness of the clustering method.

However, some papers in the literature deals with overconfidentness issue, see Park et al. (2021), in a specific context of computer vision. Aghababyan et al. (2018) proposed an entropy based metric to evaluate the confidentness of clustering algorithm without proposing any correction of the underlying probabilities in case of over/underconfident methods. Yao et al. (2000) introduced a new fuzzy clustering algorithm based on entropy without investigating about the confidentness level of the proposed algorithm with respect to other state of the art methods. In addition, to the best of author's knowledge no work has been done on proposing a statistical hypothesis test to decide about accepting or rejecting a clustering method based on its confidentness level. In this paper an entropy based clustering confidentness metric is introduced. Based on this metric one will be able to compare the performance of any two clustering algorithms and a statistical test is introduced to decide about accepting or rejecting a clustering method. Furthermore, in the context of over/underconfident clustering method, a parametric correction of the underlying probabilities is proposed in order to improve the performance of the clustering.

In this paper, we first introduce, in Section 2, a metric to measure clustering confidentness using entropy with some applications based on numeric simulations. In Section 3, a statistical test helping users to decide about accepting or rejecting a given clustering algorithm is described. Then, in Section 4, we propose a parametric correction method of the underlying probabilities of under/overconfident clustering algorithms. An application of the whole methodology on a real data set is shown in Section 5. A conclusion closes the paper.

2 Measuring clustering confidentness

Let $\mathcal{I}$ be a finite set of cluster indexes. In this work, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and for any $i \in \mathcal{I}$, X_i is a $\mathbb{R}^d$ valued random variable defined on Ω with a cumulative distribution function $F_i(\cdot)$, and a probability density function $f_i(\cdot)$ with respect to the Lebesgue measure in $\mathbb{R}^d$.

These random variables X_i, $i \in \mathcal{I}$ have distinct distributions, as they correspond to individuals of different clusters. Assume that one chooses a cluster index with a random variable I taking values in $\mathcal{I}$. One labelled observation of the corresponding individual is given by the couple (X_I, I). The purpose of a clustering procedure is to retrieve the association between observations of the mixture distribution X_I and labels I when the labels are lost.

In this paper, we aim at assessing the performance of different clustering methods. Notice that the precise value of the labels has no impact, so that a perfect clustering procedure can associate I or any permutation $\sigma(I)$ to the observations X_I. The performance measure should be insensitive to permutations of I. We present hereafter a criterion and some tests based on Shannon's entropy.

2.1 Theoretical entropy

A labeled clustering is given by the joint distribution of (X_I, I). When the labels are lost, we aim at inferring the values of the random variable I given the observations X_I.

Recall that $\mathcal{I} = \{1, 2, \ldots, |\mathcal{I}|\}$ is the finite set of all possible cluster indexes. Consider the mixture distribution given by:

$$\mathcal{G} = X_I, \text{ with } I \in \mathcal{I} \text{ and each } X_i \sim F_i,\ i \in \mathcal{I}.$$

F_i is the cumulative distribution function (*cdf*) of the random variable (*rv*) X_i.

Then, the *cdf* $F_{\mathcal{G}}$ and the probability density functions (*pdf*) $f_{\mathcal{G}}$ of $\mathcal{G}$ are as follows:

$$\begin{cases} F_{\mathcal{G}}(x) = \sum_{i \in \mathcal{I}} \alpha_i F_i(x), \\ f_{\mathcal{G}}(x) = \sum_{i \in \mathcal{I}} \alpha_i f_i(x), \end{cases}$$

where, f_i is the *pdf* of X_i, with $\alpha_i = \mathbb{P}[I = i]$ and $\sum_{i \in \mathcal{I}} \alpha_i = 1$.

Therefore, the probability that a given observation x is sampled from the underlying *rv* X_i is:

$$\begin{aligned} p_i(x) &:= \mathbb{P}[I = i \mid \mathcal{G} = x], \\ &= \frac{\alpha_i f_i(x)}{\sum_{j \in \mathcal{I}} \alpha_j f_j(x)}. \end{aligned}$$

Now, we compute Shannon's entropy of the *rv* I given $\mathcal{G} = x$, measuring the information on the fact that an observation x is sampled from the distribution

F_i, $i \in \mathcal{I}$:

$$\mathcal{H}_{\mathcal{G}}(x) = -\sum_{i \in \mathcal{I}} p_i(x) \log_2 p_i(x),$$

under the convention that $0 \log_2 0 = 0$.

Recall that I is the hidden cluster index associated to X_I. The former entropy measures the uncertainty of the clustering at a given point x. When it is equal to 0, e.g. in the case when $p_i(x) = 1$ for a given i and 0 otherwise, then it is certain that x was sampled from a specific known index i. This entropy is maximal when all $p_i(x)$ are equal, so that one has totally lost the information about the index I that has generated the observation x.
This entropy is insensitive to cluster index permutations, which is a desirable property, as stated in the introduction.

Consider the function $\mathcal{H}_{\mathcal{G}} : x \to \mathcal{H}_{\mathcal{G}}(x) \in \mathbb{R}^+$. Applied to a random argument $\mathcal{G}$, this function defines the random variable $\mathcal{H}_{\mathcal{G}}(\mathcal{G}) \in \mathbb{R}^+$. $\mathcal{H}_{\mathcal{G}}(\mathcal{G})$ measures the uncertainty of the clustering at a random point having the same distribution as $\mathcal{G}$

The distribution of this random variable shows the discriminating power of the considered clustering approach: e.g., the higher its mean, the more ambiguous the situation is. However, a mean close to zero reflects a clustering where one expects to easily associate a unique label to each point x. On the other hand, a low dispersion shows that the difficulty of cluster labeling is the same for all point x, whereas a high dispersion indicates that some points are easier to label than others.

To illustrate the function $\mathcal{H}_{\mathcal{G}}$ we consider a few basic examples:

Case A Let $\mathcal{G}$ be a Gaussian mixture distribution in one dimension with a *pdf*:

$$f_{\mathcal{G}}(x) = 0.3 f_{\mathcal{N}(0,1)}(x) + 0.5 f_{\mathcal{N}(10,1)}(x) + 0.2 f_{\mathcal{N}(3,0.1)}(x).$$

where $f_{\mathcal{N}(.,.)}$ denotes the pdf of a Gaussian r.v. with indicated parameters. The left side of Figure 1 represents the *pdf* of $\mathcal{G}$, while the right side illustrates the function $\mathcal{H}_{\mathcal{G}}$. One can remark that the labeling is perfectly accurate when the values x of $\mathcal{G}$ are far from the central area of the distribution (i.e. left and right queues of the mixture distribution). However, the difficulty of cluster labeling is higher for central values where the region is fuzzy in terms of distribution selection.

Case B Let $\mathcal{G}$ be a Dirac mixture distribution in one dimension with a probability distribution:

$$\mathrm{P}[\mathcal{G} = x] = 0.3 \mathbb{1}_{\{x=0\}} + 0.7 \mathbb{1}_{\{x=3\}}$$

where, $\mathbb{1}_{\{x=a\}}$, $a \in \mathbb{R}$ is the indicator function defined by:

$$\mathbb{1}_{\{x=a\}} = \begin{cases} 0 & \text{if } x \neq a \\ 1 & \text{if } x = a \end{cases}.$$

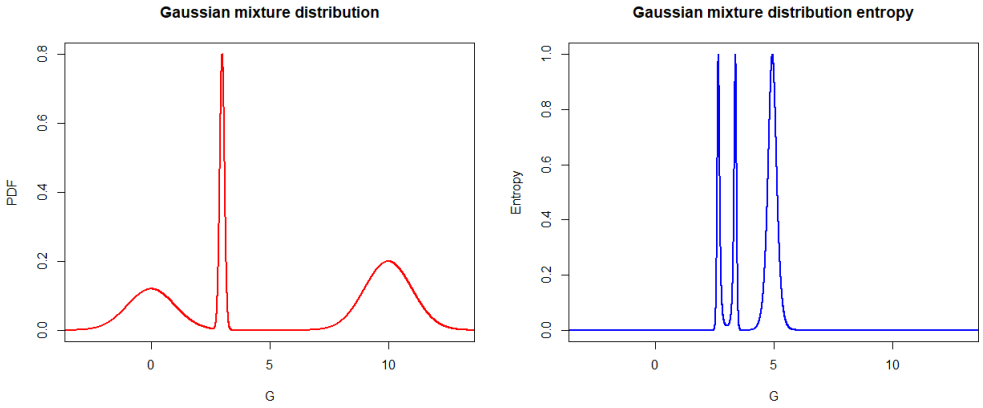

Figure 1: Gaussian mixture distribution of Case A. Left: *pdf* of the mixture distribution $\mathcal{G}$, right: entropy $\mathcal{H}_{\mathcal{G}}$ representing the variation of the labeling difficulty.

The left side of Figure 2 represents the probability masses of $\mathcal{G}$, while the right side illustrates the function $\mathcal{H}_{\mathcal{G}}$. One can remark that the labeling is perfectly accurate for all values of x. This is well suited to the Dirac case where the discrimination between different clusters/labels is obvious.

Case C Let $\mathcal{G}$ be a two dimensional Gaussian mixture distribution with a *pdf*:

$$f_{\mathcal{G}}(x) = 0.4 f_{\mathcal{N}(\mu_1,\Sigma_1)} + 0.6 f_{\mathcal{N}(\mu_2,\Sigma_2)},$$

where, $\mu_1 = (0,0)^T$ and $\mu_2 = (4,4)^T$. In addition $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\Sigma_3 = \begin{pmatrix} 2 & 0.4 \\ 0.4 & 2 \end{pmatrix}$.
The left side of Figure 3 represents the contour lines of the *pdf* of $\mathcal{G}$, while the right side illustrates the contour lines of $\mathcal{H}_{\mathcal{G}}$. Once again one can say that the accuracy of the labeling is lower on zones that are with a high fuzziness level (e.g. between the two modes), and thus difficult to label.

2.2 Empirical entropy

The idea we develop below is that a fuzzy clustering algorithm, applied to an iid sample of $\mathcal{G}$, should end up with an entropy distributed as the random variable

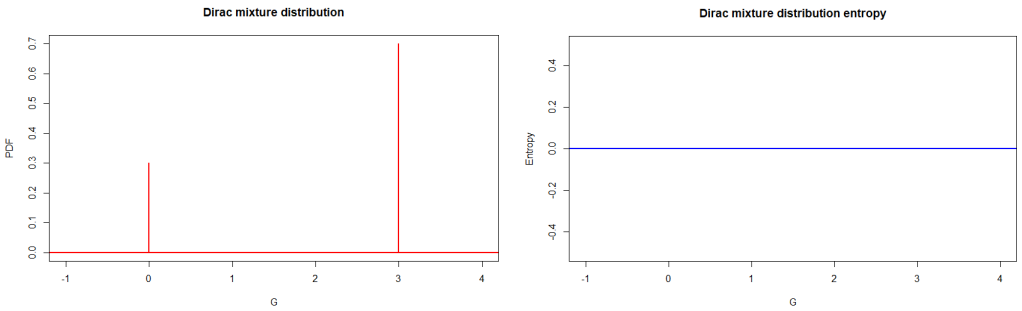

Figure 2: Dirac mixture distribution of Case B. Left: probability masses of the mixture distribution $\mathcal{G}$, right: zero-valued entropy $\mathcal{H}_{\mathcal{G}}$ representing the trivial labeling of any observation of $\mathcal{G}$.

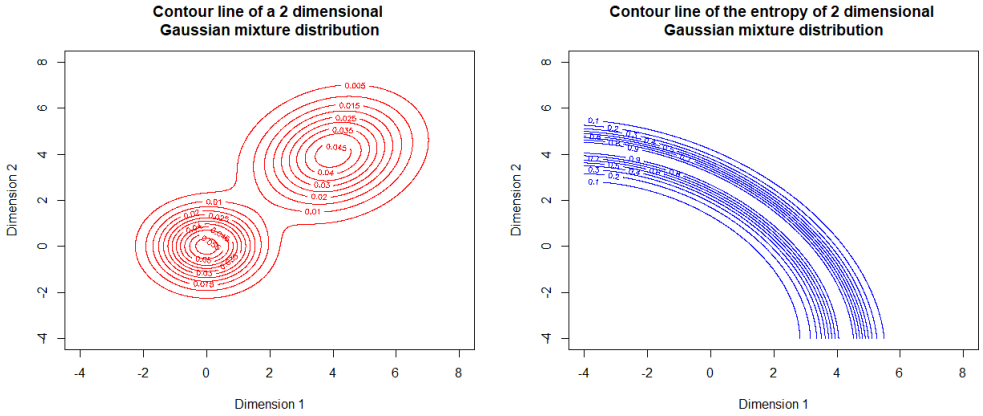

Figure 3: Two dimensional Gaussian mixture distribution of Case C Left: *pdf* of the mixture distribution $\mathcal{G}$, right: entropy $\mathcal{H}_{\mathcal{G}}$ representing the variation of the labeling difficulty.

$\mathcal{H}_{\mathcal{G}}(\mathcal{G})$. If it concludes with a lower mean entropy, then the clustering algorithm is too overconfident in its associations/labeling. If it concludes with a higher mean entropy, then the clustering algorithm is too hesitating.
Hence, the distribution of $\mathcal{H}_{\mathcal{G}}(\mathcal{G})$ reflects the difficulty of the labeling problem. In fact, depending on the underlying structure of $\mathcal{G}$, some clustering algorithms may perform better then others.

Notice that we will be able to compare algorithms that associate a cluster index probability for each point, and algorithms that associate a unique label to each point.

We start by generating n observations $\mathcal{D} = \{x_\ell\}_{\ell=1,\ldots,n}$ from an iid sample of the mixture distribution $\mathcal{G}$. Let $\mathcal{G}_n$ be a *rv* distributed as the empirical distribution of this data $\mathcal{D}$.

We run a clustering algorithm on our generated data by fixing the set of cluster indices to be a finite set $\mathcal{J}$, not necessary of the same cardinal of $\mathcal{I}$.

Then, for each generated observation x_ℓ, $\ell = 1, \ldots, n$, we get, as an output of the clustering algorithm, the probability distribution of belonging to different clusters, which will be denoted by $\left\{p_{C_j}(x_\ell)\right\}_{\substack{j\in\mathcal{J}\\ \ell=1,\ldots,n}}$, where $p_{C_j}(x_\ell)$ is the probability to associate x_ℓ to cluster c_j. Now, based on these probabilities a new Shannon's entropy distribution is computed:

$$\mathcal{H}_{\mathcal{C}}(x) = -\sum_{j\in\mathcal{J}} p_{Cj}(x) \log_2 p_{Cj}(x).$$

Considering a random argument $\mathcal{G}_n$, we aim at comparing the distributions of the two random variables $\mathcal{H}_{\mathcal{G}}(\mathcal{G})$ and $\mathcal{H}_{\mathcal{C}}(\mathcal{G}_n)$.

Therefore, in order to compare the entropy distributions of the random variables $\mathcal{H}_{\mathcal{G}}(\mathcal{G})$ and $\mathcal{H}_{\mathcal{C}}(\mathcal{G}_n)$ a distance is calculated. We denote this distance by:

$$D\left(\mathcal{H}_{\mathcal{G}}(\mathcal{G}) \,||\, \mathcal{H}_{\mathcal{C}}(\mathcal{G}_n)\right).$$

Note that, the distribution of $\mathcal{H}_{\mathcal{G}}(\mathcal{G})$ represents the theoretical labeling difficulty for the mixture distribution, whereas $\mathcal{H}_{\mathcal{C}}(\mathcal{G}_n)$ represents the perceived labeling difficulty of the clustering algorithm on the data.

Now, we reconsider the first two examples of Section 2.1 and we test the proposed approach on two clustering methods:

1. In the first application, the K-means clustering approach is applied on a sample of $n = 10\,000$ observations generated from the mixture distribution $\mathcal{G}$. In order to assess the performance of the considered method, the Kolmogorov-Smirnov test is applied as a comparison tool between the entropy distributions $\mathcal{H}_{\mathcal{G}}(\mathcal{G})$ and $\mathcal{H}_{\mathcal{C}}(\mathcal{G}_n)$. In the one dimensional Gaussian mixture model (Case A), the cardinal of $\mathcal{J}$ was considered to be equal to

three. Kolmogorov-Smirnov statistics has a value of $D\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_n\right) \,||\, \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}_n\right)\right) = 0.43$ with a $p-$value of $\approx 0\%$. Then, one can say that the two distributions are significantly different and that the proposed clustering method fails to identify the labeling difficulty. This is expected as a hard clustering here is clearly overconfident in its labeling. However, the K-means applied on Dirac mixture underlying distribution, with $\mathcal{J} = 2$, gives $D\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_n\right) \,||\, \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}_n\right)\right) \approx 0$ and a $p-$value of $\approx 100\%$. Then, in the Dirac case, K-means succeed perfectly in re-finding the original labels of the observations. Note that these results fit well with the nature of the K-means algorithm, which proposes labeling with no degree of fuzziness, known as hard clustering approach. Hence, one can conclude that such kind of clustering algorithm is to be selected when the underlying structure of $\mathcal{G}$ is more categorical, as in the case of a Dirac distribution.

2. In the second application, a soft clustering approach is applied. To do this, we have chosen the fuzzy clustering algorithm called "fanny" which was introduced in Chapter four of Kaufman and Rousseeuw (1990). Here, for each observation, instead of considering only one clustering label, probabilities of different clustering labels is computed. Now, under the same hypothesis of the first application, Kolmogorov-Smirnov statistics has a value of $\simeq 1$ with a $p-$value of 0% for both Gaussian and Dirac cases. Then, for the Gaussian underlying distribution, the considered soft approach has no additional value compared to the K-means and the algorithm fails again in identifying the labeling difficulty. However, for the Dirac case the soft approach, unlike the K-means, is not able to identify well original labels, which is logical due to the fact that in Dirac case the mass function is concentrated at only one point. Finally, the comparisons of the *cdf* of both $\mathcal{H}_{\mathcal{G}}\left(\cdot\right)$ and $\mathcal{H}_{\mathcal{C}}\left(\cdot\right)$ in Figure 4 shows that, as a result of its high fuzziness, the chosen clustering is under-confident. In fact, the values of $\mathcal{H}_{\mathcal{G}}\left(\cdot\right)$ are almost concentrated in the interval $[0, 0.15]$ while those of $\mathcal{H}_{\mathcal{C}}\left(\cdot\right)$ are spread on $[0.15, 1.4]$. In addition, Table 1 shows that, under the Gaussian assumption, the distribution of $\mathcal{H}_{\mathcal{C}}\left(\cdot\right)$ is more volatile with significantly higher mean comparing to $\mathcal{H}_{\mathcal{G}}\left(\cdot\right)$ which is another evidence of the high level of uncertainty of the soft clustering algorithm.

2.3 Jensen-Shannon Divergence

In this subsection the distance between two entropy distributions P and Q is calculated based on Jensen-Shannon divergence (JSD) metric defined by:

$$\mathrm{JSD}\left(P \,||\, Q\right) = \frac{1}{2}\left[\mathrm{KL}\left(P \,||\, R\right) + \mathrm{KL}\left(Q \,||\, R\right)\right],$$

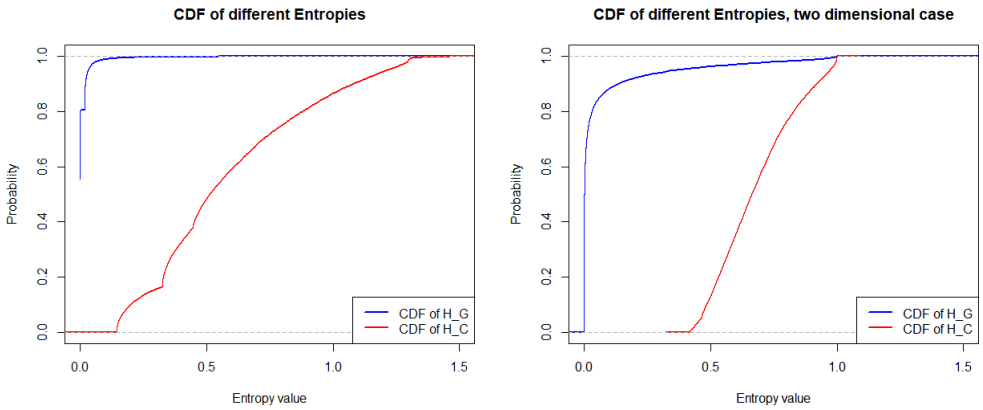

Figure 4: *CDF* of different entropies (under fanny algorithm) in the case of a Gaussian mixture distribution ($n = 10\,000$ observations). On the left: one dimensional case of Case A, on the right: two dimensional case of Case C

	1-D Gaussian case		2-D Gaussian case	
	$\mathcal{H}_{\mathcal{G}}(\mathcal{G}_n)$	$\mathcal{H}_{\mathcal{C}}(\mathcal{G}_n)$	$\mathcal{H}_{\mathcal{G}}(\mathcal{G}_n)$	$\mathcal{H}_{\mathcal{C}}(\mathcal{G}_n)$
Mean	0.008	0.585	0.056	0.681
Variance	0.0014	0.1004	0.027	0.024
Minimum	0	0.144	$\simeq 0$	0.415
Maximum	0.993	1.534	$\simeq 1$	1

Table 1: Moments of entropy distributions for soft clustering fanny-algorithm ($n = 10\,000$ observations)

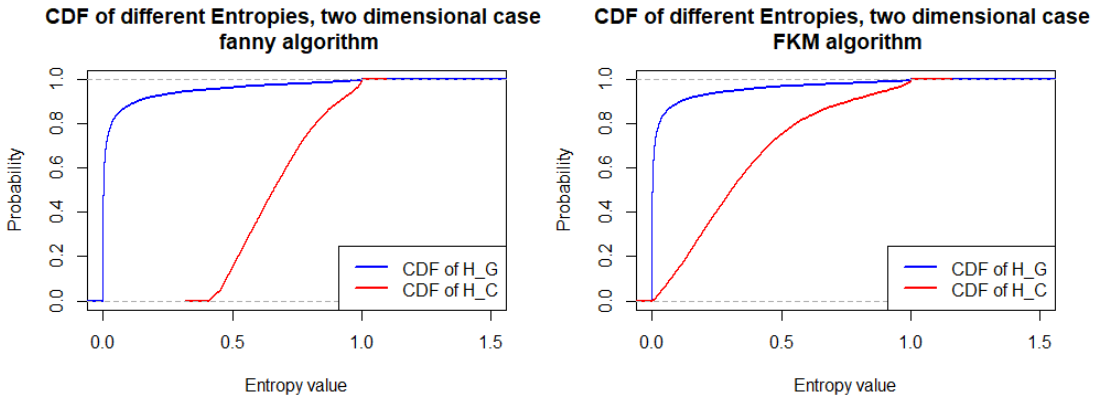

Figure 5: CDF of different entropies in the case of a 2-D Gaussian mixture model, Case C, ($n = 10\,000$ observations) for fanny and FKM algorithms.

where $R = \frac{1}{2}(P + Q)$ and $\mathrm{KL}(P \,||\, R)$ denotes the Kullback-Leibler divergence between probability distributions P and R. Note that the JSD has the following properties: 1) Non negative measure; 2) Symmetric measure and 3) JSD $\in [0, 1]$, with JSD $= 0$ if and only if $P = Q$. In addition, this metric is more suited to the case of fuzzy underlying distributions when soft clustering algorithm are applied.

Then, by applying the JSD on our previous examples we get the following results:

1. For the Gaussian mixture distribution and "fanny" clustering algorithm context with a sample size $n = 10\,000$: $\mathrm{JSD}\left(\mathcal{H}_{\mathcal{G}}(\mathcal{G}_n) \,||\, \mathcal{H}_{\mathcal{C}}(\mathcal{G}_n)\right) = 0.985$ for one dimensional Gaussian case and $\mathrm{JSD}\left(\mathcal{H}_{\mathcal{G}}(\mathcal{G}_n) \,||\, \mathcal{H}_{\mathcal{C}}(\mathcal{G}_n)\right) = 0.897$ for the two dimensional case. Then, one can say that the performance of "fanny" algorithm is better in two dimensional Gaussian mixture distribution but it still an under-confident approach (see Figure 4 and Table 1). Note that, in the numerical illustrations of this paper, in order to compute different JSD, random variables $\mathcal{H}_{\mathcal{G}}(\mathcal{G}_n)$ and $\mathcal{H}_{\mathcal{C}}(\mathcal{G}_n)$ were discretized with a constant step size of 0.001.

2. Now in order to compare the efficiency of two different soft clustering algorithm "fanny" and "Fuzzy K-means (FKM)" introduced by Bezdek (1981) we consider a sample $n = 10\,000$ of the two dimensional Gaussian mixture distribution Case C introduced in Section 2.1. Under these assumptions we get $\mathrm{JSD}\left(\mathcal{H}_{\mathcal{G}}(\mathcal{G}_n) \,||\, \mathcal{H}_{\mathcal{C}}(\mathcal{G}_n)\right)^{\mathrm{fanny}} = 0.897 > \mathrm{JSD}\left(\mathcal{H}_{\mathcal{G}}(\mathcal{G}_n) \,||\, \mathcal{H}_{\mathcal{C}}(\mathcal{G}_n)\right)^{\mathrm{FKM}} = 0.658$. Then the FKM algorithm is more suited to find the hidden labels of such data. The result is illustrated in Figure 5.

n	1-D Gaussian case	2-D Gaussian case
100	1.000	1.000
500	0.991	0.967
1000	0.997	0.959
3000	0.990	0.921
5000	0.990	0.920

(a) Performance of "fanny" algorithm for different values of n

n	fanny	FKM
100	1.000	0.960
500	0.967	0.846
1000	0.959	0.817
3000	0.921	0.707
5000	0.920	0.664

(b) Comparing the performance of "fanny" and FKM algorithms for different values of n, Case C

Table 2: JSD for different values of n, the lower the better.

To study the impact of the sample size n on the quality and the performance of different clustering algorithms, Table 2a shows $\mathrm{JSD}\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_n\right) \,||\, \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}_n\right)\right)$ for different values of n in the case of one and two dimensional Gaussian underlying distribution and in the context of "fanny" clustering algorithm. On the other hand Table 2b shows a comparison between $\mathrm{JSD}\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_n\right) \,||\, \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}_n\right)\right)^{\mathrm{fanny}}$ and $\mathrm{JSD}\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_n\right) \,||\, \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}_n\right)\right)^{\mathrm{FKM}}$ for different values of n in the case of two dimensional Gaussian mixture distribution. In both cases, especially for 2-D Gaussian case, JSD is decreasing for high values of n. Then, one can say that the ability of a clustering algorithm to reconstruct original hidden labels is partly related to the uncertainty on the empirical distribution.

As not to be limited to the comparative aspect of our approach, its important to propose a method that helps setting a kind of threshold or boundary between accepted and rejected clustering method, in terms of labels reconstruction performance, for a given set of data.

3 Statistical test

Using the results of Section 2, we can also make some progress on the problem of selecting a good clustering method that fits well the original mixture distribution assumption. We consider the context of a statistical test with H_0 : "the entropy of clustering method fits the original mixture distribution assumption" vs H_1 : "the entropy of the clustering method doesn't fit the mixture distribution assumption". In other words one can write:

$$H_0 : \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}\right) \overset{d}{=} \mathcal{H}_{\mathcal{G}}\left(\mathcal{G}\right).$$

To this end, the problem of obtaining an approximate critical region of the considered test, under H_0, can be solved by applying a numerical simulation method. To specify this critical region, two samples of n_1 and n_2 observations respectively are generated according to the mixture distribution $\mathcal{G}$. The distance between $\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_{n_1}\right)$ and $\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_{n_2}\right)$ is calculated. By repeating this procedure a large number of times, one obtains the approximate quantiles of the

distribution of $D\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_{n_1}\right) \,||\, \mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_{n_2}\right)\right)$ from which one can get a critical region of approximated level α. Then, a clustering method, on a sample of n observations, is accepted if $D\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_{n}\right) \,||\, \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}_{n}\right)\right) < (1-\alpha)^{th}$ quantile of the obtained distance distribution.

If Kolmogorov Smirnov statistic is used to compute the distance between $\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}\right)$ and $\mathcal{H}_{\mathcal{C}}\left(\mathcal{G}\right)$, then the decision of accepting or rejecting H_0 can be directly assessed based on the quantile of a Kolmogorov distribution under a given asymptotic confidence level α.

As an application, using JSD metric, we consider testing H_0 vs H_1 where the underlying distribution is the two dimensional Gaussian mixture defined in Section 2.1. Based on the numerical simulation method described above, with $n_1 = n_2 = 10\,000$, and by repeating the procedure 1000 times, we obtain the approximate quantiles of the corresponding JSD distribution and the critical region of approximated level $\alpha = 5\%$, which is $q_{0.95}^{JSD} = 0.04$. Then for a given clustering algorithm if $\text{JSD}\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_{n}\right) \,||\, \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}_{n}\right)\right) > q_{0.95}^{JSD}$ we reject H_0. Hence the clustering method fails to understand the labeling difficulty of the considered observation. In our case, both fanny and FKM are rejected as both $\text{JSD}\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_{n}\right) \,||\, \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}_{n}\right)\right)^{\text{fanny}}$ and $\text{JSD}\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_{n}\right) \,||\, \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}_{n}\right)\right)^{\text{FKM}}$ are greater than 0.04.

4 Correction of clustering probabilities

Assume that the statistical test of Section 3 is rejecting the clustering method. Before proposing another clustering algorithm, a possible solution is to adjust the underlying cluster probabilities using a parametric transformation.

Recall that $p_{C_j}\left(x_\ell\right)$ is the probability to affect observation x_ℓ to cluster c_j. In order to adjust these probabilities to fit the original mixture distribution we propose the following transformation:

$$p_{C_j}\left(x_\ell\right)^* = \frac{p_{C_j}\left(x_\ell\right)^\theta}{\sum_{i\in\mathcal{J}} p_{C_i}\left(x_\ell\right)^\theta}, \tag{1}$$

where $\mathcal{J}$ is the proposed set of cluster indices and $\theta \in \mathbb{R}^+$ a parameter, to be estimated, reflecting the reason behind the bad clustering quality.

Then, based on these new probabilities a new clustering entropy is calculated. We denote it by $\mathcal{H}_{\mathcal{C}}\left(\mathcal{G}, \theta\right)$. Hence, one may assess the effect of this transformation by recomputing $\text{JSD}\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_{n}\right) \,||\, \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}_{n}, \theta\right)\right)$.

Now, the optimal value of the parameter θ is the one minimizing the previous JSD, which will be given by:

$$\hat{\theta} = \arg\min_{\theta} \text{JSD}\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_{n}\right) \,||\, \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}_{n}, \theta\right)\right).$$

So, in case $\hat{\theta} > 1$, one can say that the proposed clustering algorithm is not categorical enough and has a higher fuzziness level then the one of the original

	Training set			Testing set	
	$\text{JSD}(\theta = 1)$	$\hat{\theta}$	$\text{JSD}\left(\hat{\theta}\right)$	$\text{JSD}(\theta = 1)$	$\text{JSD}\left(\hat{\theta}\right)$
Fanny	0.895	5.958	0.043	0.918	0.088
FKM	0.630	3.348	0.039	0.678	0.084

Table 3: Comparison of JSD: original vs corrected on testing and training data

mixture distribution. Otherwise, if $\hat{\theta} < 1$, the current clustering algorithm is too categorical and we need to regularize it by injecting some source of fuzziness. Due to this interpretation, the parameter $\hat{\theta}$ can be considered as a fuzziness index.

To study the efficiency of this approach, we apply it on a previous example considering two different soft clustering algorithm "fanny" and "Fuzzy K-means (FKM)". We generate a sample $n = 10\,000$ of the two dimensional Gaussian mixture distribution with the same parameters as the one in Section 2.1. Based on this training sample we calculate the estimator $\hat{\theta}$ of θ. Then, we use the value of this estimator to correct the probabilities proposed by each of the clustering algorithm on a testing sample of 3000 observations and we compare the impact of this correction on the quality of the clustering on both training and testing sets by calculating the corresponding JSD. Under these assumptions we get results summarized in Table 3.

Then, by applying the parametric transformation proposed in Equation (1) we are able to reduce enormously the JSD and correcting the behavior of both algorithms to be more suited to find the hidden labels of the original data. Even, on the testing set the correction has a significant positive impact on the clustering quality for both algorithm. In addition, comparison of the *cdf* of different Entropies $\left(\mathcal{H}_{\mathcal{G}}\left(\mathcal{G}_n\right) \text{ vs } \mathcal{H}_{\mathcal{C}}\left(\mathcal{G}_n, \hat{\theta}\right)\right)$ are illustrated in Figure 6 for training set and Figure 7 for testing.

Note that a more complex transformation than the one proposed in (1) can be done using more than one parameter and focusing on local parts of the distribution. By adapting the same approach described in this section we can select the best transformation according to, say, the least JSD criterion. Also, as the application is done on data with gaussian underlying distributions, it is obvious to expect that the best performance, in terms of clustering, will be for the Gaussian Mixture Model (GMM) based clustering method, introduced by McLachlan and Basford (1988), where gaussian distributions are considered as priors. As expected, even before correction, JSD of the GMM method is about 0.042 for training and 0.089 for testing with a good fitting of the CDFs of the two entropies (see Figure 8).

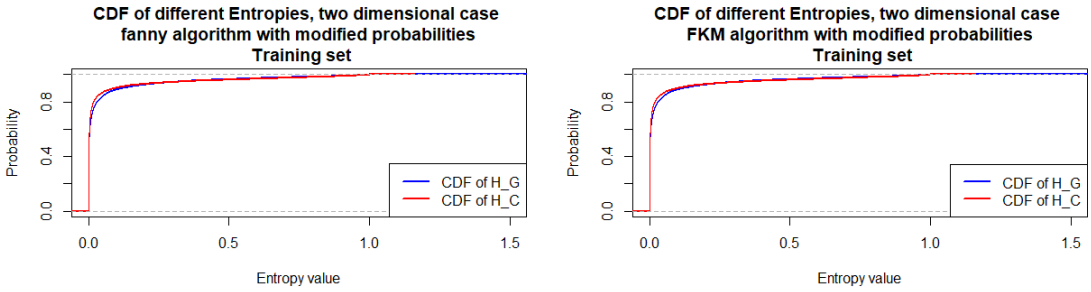

Figure 6: CDF of different entropies in the case of a 2-D Gaussian mixture model ($n = 10\,000$ observations, Case C) for fanny ($\hat{\theta} = 5.958$) and FKM algorithms ($\hat{\theta} = 3.348$) with modified probabilities applied on training data

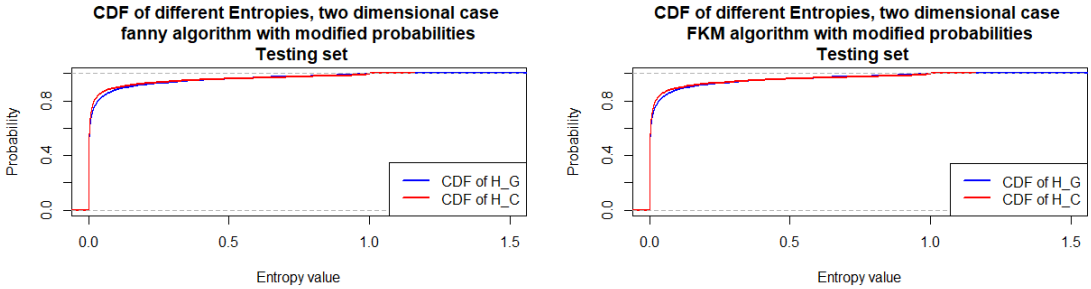

Figure 7: CDF of different entropies in the case of a 2-D Gaussian mixture model ($n = 3\,000$ observations, Case C) for fanny ($\hat{\theta} = 5.958$) and FKM algorithms ($\hat{\theta} = 3.348$) with modified probabilities applied on testing data

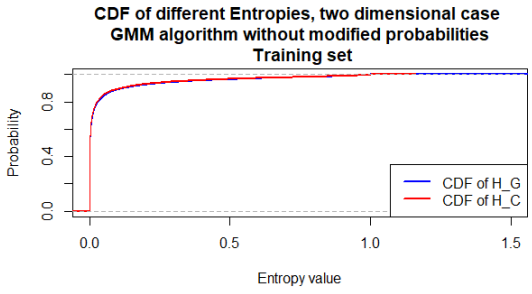

Figure 8: CDF of different entropies in the case of a 2-D Gaussian mixture model ($n = 10\,000$ observations, Case C) for GMM algorithm ($\hat{\theta} = 0.970$)

Furthermore, in order to assess the ability of fuzzy clustering methods in reconstructing the original clusters of an underlying data, after the application of the proposed probability transformation, we used the theoretical labels of the 2-D Gaussian mixture distribution generated in Subsection 2.1. Then, the idea is to make a confusion matrix comparing theoretical labels with the labels proposed by the clustering method. The accuracy of each clustering method is calculated as the proportion of well classified observation and formulated by:

$$Accuracy = \frac{1}{n}\sum_{\ell=1}^{n}\sum_{j\in\mathcal{J}}\mathbb{P}\left[\text{assign observation } \ell \text{ to cluster } j / \text{Theoretically observation } \ell \text{ is in cluster } j\right]. \tag{2}$$

Tables 4 and 5 show the clustering accuracy as defined in Equation (2) for several soft clustering method applied on 2-D Gaussian mixture distribution generated data, before and after applying the correction of clustering probabilities on training and testing sets respectively. Note that here also θ is estimated on the training set and used to correct clustering probabilities on the testing set. In general, the accuracy of the clustering is improving in a significant way after applying the correction method, except for the GMM clustering method where the accuracy is almost the same as this method is designed to deal with gaussian mixture underlying distributions which is fitting perfectly the generated data. In addition, the estimated value of θ is giving information about the quality of the clustering method. For Fanny and FKM algorithms the fuzziness index $\hat{\theta}$ is far greater than one, so these two methods are not categorical enough and have higher fuzziness level than the original mixture. However, $\hat{\theta}$ for GMM is close to one, which is reasonable because GMM is designed in a way to generate original clusters of a gaussian mixture distribution. Also, an added value of the proposed correction method, is that after performing the correction one can revisit the result of the statistical test that was introduced in Section 3 to reassess if after modification a certain clustering method, that was rejected previously, is now accepted. Conclusions of the statistical test applied on several soft clustering methods before and after probability correction are presented in Tables 6 and 7 for training and testing sets respectively. It is clear that after applying the parametric correction of the probabilities, the values of the statistic of the test of Section 3 decrease significantly to become much closer to the critical region empirical quantile $\left(q_{0.95}^{JSD} = 0.04\right)$. In fact some of the clustering algorithms that was rejected before correction are accepted after it. Then, correcting clustering probabilities makes: 1) different clustering method more accurate in reconstructing original clusters and 2) the comparison between different clustering method more reasonable in addition to the selection of the best performing one.

Clustering method	Accuracy before correction	Accuracy after correction	$\hat{\theta}$
GMM	0.968	0.980	0.970
Fanny	0.802	0.968	5.958
FKM	0.904	0.969	3.348

Table 4: Accuracy of different clustering methods calculated on training data before and after correction, the higher the better. θ is estimated based on the training set and used to correct clustering probabilities on both training and testing sets.

Clustering method	Accuracy before correction	Accuracy after correction	$\hat{\theta}$
GMM	0.978	0.978	0.970
Fanny	0.798	0.967	5.958
FKM	0.901	0.967	3.348

Table 5: Accuracy of different clustering methods calculated on testing data before and after correction, the higher the better. θ is estimated based on the training set and used to correct clustering probabilities on both training and testing sets.

	Value of the JSD statistic		Decision about H_0	
Clustering method	Before correction	After correction	Before correction	After correction
GMM	0.042	0.035	Accepted	Accepted
Fanny	0.895	0.043	Rejected	Accepted
FKM	0.630	0.039	Rejected	Accepted

Table 6: Statistical test conclusions on training data. JSD is calculated on training data before and after correction, the lower the better. θ is estimated based on the training set and used to correct clustering probabilities on both training and testing sets.

	Value of the JSD statistic		Decision about H_0	
Clustering method	Before correction	After correction	Before correction	After correction
GMM	0.089	0.086	Rejected	Rejected
Fanny	0.918	0.088	Rejected	Rejected
FKM	0.678	0.084	Rejected	Rejected

Table 7: Statistical test conclusions on testing data. JSD is calculated on testing data before and after correction, the lower the better. θ is estimated based on the training set and used to correct clustering probabilities on both training and testing sets.

5 Application on real data

As an application of our approach, we use an open source data from the UCI Machine Learning Repository, Dua and Graff (2017), available at `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29`. This data classifies $n = 569$ individuals in either malignant (M) or benign (B) breast cancer using 30 numeric features characterizing each person. I.e., a space of dimension $\mathbb{R}^{569\times 30}$, labeled into $\mathcal{I} = \{1 = \text{"M"}, 2 = \text{"B"}\}$ cluster indexes. Therefore, assuming that we are in the context of a mixture distribution $\mathcal{G}$, the probability that a given observation x is sampled from the underlying rv X_i can be estimated for any $i \in \mathcal{I}$, by for example:

$$\widehat{p_i}(x) = \frac{\widehat{\alpha}_i \widehat{f^i}(x)}{\sum_{j\in\mathcal{I}} \widehat{\alpha}_j \widehat{f^j}(x)}, \tag{3}$$

with, $\widehat{f^i}(\cdot)$ the kernel density estimation (KDE) of the observations generated by the underlying rv X_i (i.e., observations in the cluster of index i). In other words it is an estimation of the pdf of the rv X_i. . Note that, practically one can use kernel density estimation (KDE) functions, already implemented in statistical software, to compute these values. The quantity $\widehat{\alpha}_i$ is the proportion of observations in the same cluster, i.e.,

$$\widehat{\alpha}_i = \frac{\text{Number of observations in cluster } i}{n}.$$

For the three different soft clustering methods introduced previously, we compare the impact of the parametric correction proposed in Equation (1) on the quality of the clustering by first estimating the optimal value of θ, on a training set representing 80% of the considered data set, and then calculating the corresponding JSD on both training and testing sets. Results are in Tables 8 and 9.

Considering clustering fuzziness, Fanny is the worst clustering method for the considered data, even after the correction it is affecting the same probability for both clusters (1 and 2) without any discrimination power. On the other hand, the performance of FKM and GMM is improving enormously after the correction with a preference to the GMM method. The significant impact of the parametric correction on FKM and GMM can be seen also when comparing the CDF of the entropies before and after correction in Figures 9 and 10. Note that, the impact of the correction for the GMM is limited because even before correction GMM is performing well on reconstructing original labels. In addition, one can remark that the estimated value of θ for the FKM is far higher than one which is equivalent to say that FKM has a higher fuzziness level than the original mixture of the considered cancer data set. However,

	Training set		
	JSD($\theta = 1$)	$\hat{\theta}$	JSD$\left(\hat{\theta}\right)$
Fanny	1	0.001	1
FKM	0.989	9.883	0.249
GMM	0.0834	1.652	0.0556

Table 8: Comparison of JSD: original vs corrected on training data. θ is estimated on the training set and used to compute JSD (the lower the better) on both training and testing sets.

	Testing set		
	$\text{JSD}(\theta = 1)$	$\hat{\theta}$	$\text{JSD}\left(\hat{\theta}\right)$
Fanny	1	0.001	1
FKM	1	9.883	0.352
GMM	0.0709	1.652	0.0827

Table 9: Comparison of JSD: original vs corrected on testing data. θ is estimated on the training set and used to compute JSD (the lower the better) on both training and testing sets.

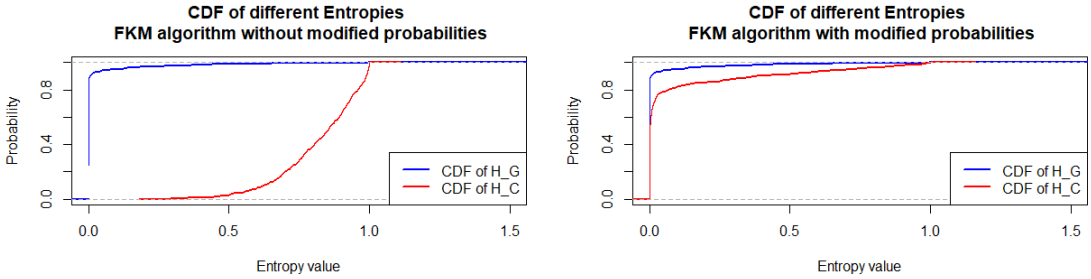

Figure 9: CDF of different entropies for FKM algorithms with and without modified probabilities applied on cancer data

the fuzziness index $\hat{\theta}$ for GMM is close to one, which is reasonable because GMM is already performing well on discrimination level even before the correction.

Now, by focusing only on FKM and GMM we can also compute the accuracy of each clustering method before and after the parametric correction based on the approach described in Equation (2). Here also $\hat{\theta}$ is estimated on the training set and used to compute the clustering accuracy for both training and testing sets. Results are in Tables 10 and 11. On both training and testing datasets one may remark that the clustering accuracy of the FKM method is improving significantly after the correction of the corresponding probabilities. However, the correction has no impact on the accuracy of the GMM. This is also reflected by the value of $\hat{\theta}$ which is close to one, indicating that the confidentness level of the GMM is acceptable and is capturing well the original labels of the clusters. This may be explained by the possibility that the underlying distribution of the considered dataset is following a gaussian mixture distribution.

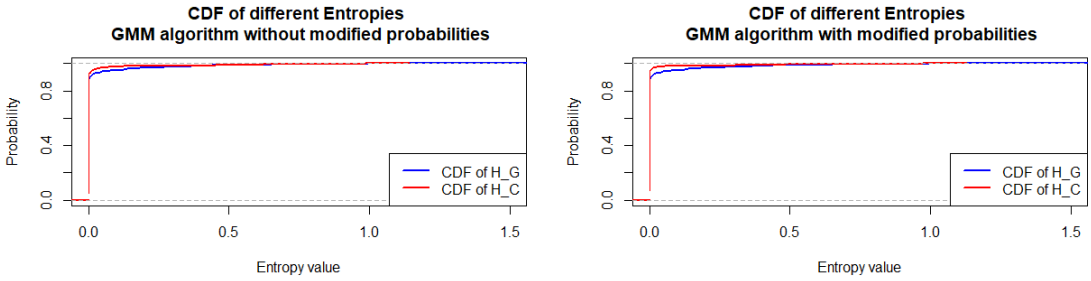

Figure 10: CDF of different entropies for GMM algorithms with and without modified probabilities applied on cancer data

Clustering method	Accuracy before correction	Accuracy after correction	$\hat{\theta}$
FKM	0.7	0.91	9.883
GMM	0.937	0.937	1.652

Table 10: Accuracy of different clustering methods based on training data, the higher the better. θ is estimated on the training set and used to compute JSD on both training and testing sets.

Clustering method	Accuracy before correction	Accuracy after correction	$\hat{\theta}$
FKM	0.683	0.89	9.883
GMM	0.957	0.955	1.652

Table 11: Accuracy of different clustering methods based on testing data, the higher the better. θ is estimated on the training set and used to compute JSD on both training and testing sets.

6 Conclusion

This paper introduced an innovative metric, based on Shannon's entropy, assessing the quality of clustering algorithms in term of confidentness level . The proposed metric can be used to compare the performance of two clustering algorithm in a way to conclude which is more over/under confident. In addition a statistical test, based also on the introduced metric, has been constructed to help taking the decision about accepting or rejecting a clustering algorithm using its confidentness quality. Moreover, a parametric adjustment of the underlying probabilities of a clustering algorithm has been introduced in order to improve the confidentness quality of the corresponding clustering method. According to many numerical simulations and real world data application it was noticed that the proposed methodology helps users significantly in getting better discrimination power from a given clustering method. So far, a first perspective of this work is to try to develop a theoretical proof of the probability distribution of the proposed metric. A second perspective is to adapt the proposed methodology on data of higher level of complexity (i.e., data of mixed typology). Finally, this work may be developed in a way to use the different metrics to identify the optimal number of clusters when applying a particular clustering algorithm.

References

Aghababyan, A., Lewkow, N., and Baker, R. S. (2018). Enhancing the clustering of student performance using the variation in confidence. In *International Conference on Intelligent Tutoring Systems*, pages 274–279. Springer.

Bataineh, K. M., Naji, M., and Saqer, M. (2011). A comparison study between various fuzzy clustering algorithms. *Jordan Journal of Mechanical & Industrial Engineering*, 5(4).

Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms: Advanced applications in pattern recognition.

Dawyndt, P., Meyer, H. D., and Baets, B. D. (2005). The complete linkage clustering algorithm revisited. *Soft Computing*, 9(5):385–392.

De Oliveira, J. V. and Pedrycz, W. (2007). *Advances in fuzzy clustering and its applications*. John Wiley & Sons.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., and Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743.

Forrest, S. (1996). Genetic algorithms. *ACM Computing Surveys (CSUR)*, 28(1):77–80.

Kaufman, L. and Rousseeuw, P. J. (1990). *An introduction to cluster analysis*. John Wiley and Sons, Incorporated.

Kriegel, H.-P., Kröger, P., Sander, J., and Zimek, A. (2011). Density-based clustering. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(3):231–240.

MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York.

Park, S., Han, S., Kim, S., Kim, D., Park, S., Hong, S., and Cha, M. (2021). Improving unsupervised image clustering with robust learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12278–12287.

Roux, M. (2015). A comparative study of divisive hierarchical clustering algorithms. *arXiv preprint arXiv:1506.08977*.

Ruspini, E. H., Bezdek, J. C., and Keller, J. M. (2019). Fuzzy clustering: A historical perspective. *IEEE Computational Intelligence Magazine*, 14(1):45–55.

Yang, M.-S. (1993). A survey of fuzzy clustering. *Mathematical and Computer modelling*, 18(11):1–16.

Yao, J., Dash, M., Tan, S., and Liu, H. (2000). Entropy-based fuzzy clustering and fuzzy modeling. *Fuzzy sets and Systems*, 113(3):381–388.

Contents