Neurocomputing 74.16 (2011): 2657-2664

DOI: http://dx.doi.org/10.1016/j.neucom.2011.03.023

The effect of low number of points in clustering validation via the negentropy increment

Luis F. Lago-Fernández[*,a], Manuel Sánchez-Montañés[a], Fernando Corbacho[b]

[a] *Departamento de Ingeniería Informática, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Madrid, Spain*
[b] *Cognodata Consulting, Calle Caracas 23, 28010 Madrid, Spain*

Abstract

We recently introduced the *negentropy increment*, a validity index for crisp clustering that quantifies the average normality of the clustering partitions using the negentropy. This index can satisfactorily deal with clusters with heterogeneous orientations, scales and densities. One of the main advantages of the index is the simplicity of its calculation, which only requires the computation of the log-determinants of the covariance matrices and the prior probabilities of each cluster. The negentropy increment provides validation results which are in general better than those from other classic cluster validity indices. However, when the number of data points in a partition region is small, the quality in the estimation of the log-determinant of the covariance matrix can be very poor. This affects the proper quantification of the index and therefore the quality of the clustering, so additional requirements such as limitations on the minimum number of points in each region are needed. Although this kind of constraints can provide good results, they need to be adjusted depending on parameters such as the dimension of the data space. In this article we investigate how the estimation of the negentropy increment of a clustering partition is affected by the presence of regions with small number of points. We find that the error in this estimation depends on the number of points in each region, but not on the scale or orientation of their

[*]Corresponding author. Tel.: +34 91 497 22 11; fax: +34 91 497 22 35.
Email addresses: luis.lago@uam.es (Luis F. Lago-Fernández), manuel.smontanes@uam.es (Manuel Sánchez-Montañés), fernando.corbacho@cognodata.com (Fernando Corbacho)

distribution, and show how to correct this error in order to obtain an unbiased estimator of the negentropy increment. We also quantify the amount of uncertainty in the estimation. As we show, both for 2D synthetic problems and multidimensional real benchmark problems, these results can be used to validate clustering partitions with a substantial improvement.

Key words: Crisp clustering, Cluster validation, Negentropy increment.

---

1. Introduction

The goal of cluster analysis [12] is the automatic partition of a data set into a finite number of natural structures, or clusters, so that the elements inside a cluster are similar while those belonging to different clusters are not. Clustering algorithms are usually divided into crisp and fuzzy. In crisp clustering, each data point is uniquely assigned to a single cluster. On the contrary, fuzzy clustering allows each point to belong to any of the clusters with a certain degree of membership. A common problem of both approaches is the lack of a general framework to measure the validity of the outcomes of a particular clustering method, and in particular to assess the correct number of clusters. This is the subject of cluster validation [15], whose objective is to provide a quality measure, or validity index, that allows to evaluate the results obtained by a clustering algorithm.

Many cluster validity indices have been proposed in the literature both for crisp and fuzzy clustering, including geometric [8, 2, 11, 24, 4], probabilistic [5, 3, 14], graph theoretic [25], and visual [16, 10] approaches. The concept of cluster validation is also related to the determination of the number of components in mixture models [13, 26, 22, 27].

In recent work [18, 19] we proposed a new validity index for crisp clustering that is based on the normality of the clusters. The normal or Gaussian distribution maximizes the entropy for a fixed covariance matrix [7]. This means that a normally distributed cluster has the lowest possible structure, and so it should not be further partitioned into additional substructures. This principle suggests to select the partition for which the clusters are on average more Gaussian. The normality of a probability distribution can be measured by means of the negentropy, which is defined as the difference between the entropy of the distribution and the entropy of a Gaussian with the same covariance matrix. The negentropy is a frequently used measure of

distance to normality [6], and has been widely used to measure normality in the context of projection pursuit and independent component analysis [17].

The lower the negentropy of the cluster, the more Gaussian it is. So, according to the previous principle, cluster validation should be accomplished by selecting the partition whose clusters have on average lower negentropy. The most involved technical issue regarding the use of the negentropy to measure the normality of a partition is related to the computation of the differential entropy of the clusters. However it can be shown that, by subtracting the negentropy of the initial data distribution (no partition), one obtains a validity index that: (i) measures the increment in normality that is gained with the partition; and (ii) avoids the explicit computation of any differential entropies. This index is called the *negentropy increment* [18, 19] associated to the partition.

The negentropy increment provides good results as a validity index for crisp clustering partitions when compared to other indices in the literature. In [18] it was evaluated on a set of randomly generated problems in 2D, and it showed a good performance with respect to the assessment of the number of clusters. It has also been tested on a more extensive set of problems, including real benchmark databases, where in general it provides better results than other validity indices, both with respect to the assessment of the number of clusters and to the similarity amongst the real clusters and the selected partitions [19]. Although this index can satisfactorily manage clusters with heterogeneous orientations, scales and densities, the correct evaluation of the negentropy increment is difficult in cases where some of the partition regions have only a few number of points. In such situations the negentropy increment tends to be underestimated, which leads to the selection of low quality partitions that overestimate the number of clusters. Although this effect can be alleviated by limiting the minimum number of points in each region, the adjustment of this number must be done ad hoc and depends on parameters such as the dimension of the data space.

In this article we present an extension of our previous work [18, 19] that addresses this problem. We investigate how the estimation of the negentropy increment of a clustering partition is affected by the presence of regions with only a few points. We show that the error in this estimation depends only on the number of points in each region, but not on the scale or orientation of their distribution. We introduce a correction term that eliminates the bias in the estimator of the negentropy increment, and quantify the amount of uncertainty in the estimation. Then we show how this can be used to validate

clustering partitions with substantial improvement.

The rest of the article is organized as follows. In section 2 we describe the negentropy increment of a clustering partition. In section 3 we give an overview of our previous results regarding the use of the negentropy increment as a cluster validity index. In section 4 we analyze how the presence of partition regions with low number of points affects the evaluation of the index. Section 5 introduces the correction term that eliminates the bias in the estimator of the negentropy increment. Section 6 describes the algorithm used to optimize the corrected estimator, and in section 7 we present the results on simulated and real datasets. Finally, in section 8 we present the conclusions.

2. The negentropy increment of a partition

The negentropy of a continuous random variable $\mathbf{x}$ is defined as:

$$J(\mathbf{x}) = \hat{H}(\mathbf{x}) - H(\mathbf{x}) \tag{1}$$

where $H(\mathbf{x})$ is the differential entropy of $\mathbf{x}$ and $\hat{H}(\mathbf{x})$ is the differential entropy of a normal distribution with the same covariance matrix as $\mathbf{x}$. Due to the maximum entropy property of the normal distribution [7], the negentropy is always equal to or greater than 0, with equality holding if and only if $\mathbf{x}$ is normally distributed.

Let $P = \{\Omega_1, \Omega_2, ..., \Omega_k\}$ be a crisp partition of the space into a set of k non overlapping regions, and let us consider the average negentropy of $\mathbf{x}$ across regions, $\bar{J}(\mathbf{x})$:

$$\bar{J}(\mathbf{x}) = \sum_{i=1}^{k} p_i J_i(\mathbf{x}) \tag{2}$$

where p_i is the probability of $\mathbf{x}$ falling into the region Ω_i, and $J_i(\mathbf{x})$ is the negentropy of $\mathbf{x}$ in the region Ω_i. $\bar{J}(\mathbf{x})$ is a measure of the average distance to normality of the distribution of $\mathbf{x}$ in each region. Note that any constant can be added to $\bar{J}(\mathbf{x})$ so, instead of equation 2, it is possible to consider the index:

$$\Delta J = \bar{J}(\mathbf{x}) - J_0(\mathbf{x}) \tag{3}$$

where $J_0(\mathbf{x})$ is the negentropy of $\mathbf{x}$ when no partition is performed, which is a constant for each problem. This index is called the *negentropy increment*

of the partition [18, 19], and it measures the change in negentropy due to performing a partition on the space.

The main advantage of computing the difference of negentropies in equation 3 is that, after some manipulations, the differential entropies cancel out and the negentropy increment can be written in a very simple manner [19]:

$$\Delta J = \frac{1}{2}\sum_{i=1}^{k} p_i \log|\Sigma_i| - \frac{1}{2}\log|\Sigma_0| - \sum_{i=1}^{k} p_i \log p_i \tag{4}$$

where Σ_0 is the covariance matrix of $\mathbf{x}$ and Σ_i is the covariance matrix of $\mathbf{x}$ restricted to the region Ω_i. Note that in order to evaluate this final expression we only need to compute the probabilities p_i and the log-determinants of the covariance matrices in each region.

3. The negentropy increment as a cluster validity index

The negentropy increment can be applied as a general tool to validate the outcome of a crisp clustering algorithm, and also to compare solutions provided by different algorithms for a single problem when normality is a desired property of the clusters. Given two different partitions of the data, the one with lower ΔJ will have clusters that are on average more Gaussian, and so will be preferred. In practical terms, the negentropy increment of the partition $P = \{\Omega_1, \Omega_2, ..., \Omega_k\}$ is estimated by:

$$\Delta J_B(P) = \frac{1}{2}\sum_{i=1}^{k} \tilde{p}_i \log|\tilde{\Sigma}_i| - \frac{1}{2}\log|\tilde{\Sigma}_0| - \sum_{i=1}^{k} \tilde{p}_i \log \tilde{p}_i \tag{5}$$

where $\tilde{\Sigma}_i$ are the estimated covariance matrices:

$$\tilde{\Sigma}_i = \frac{1}{N_i - 1}\sum_{\mathbf{x}\in\Omega_i} (\mathbf{x} - \tilde{\boldsymbol{\mu}}_i)\cdot(\mathbf{x} - \tilde{\boldsymbol{\mu}}_i)^T \tag{6}$$

and $\tilde{p}_i = N_i/N$. Here N_i is the number of data points in Ω_i, N is the total number of points in the problem, and $\tilde{\boldsymbol{\mu}}_i$ is the sample mean for the region Ω_i, $\tilde{\boldsymbol{\mu}}_i = \frac{1}{N_i}\sum_{\mathbf{x}\in\Omega_i} \mathbf{x}$. As we will see, the expression in equation 5 is a biased estimator of ΔJ. For the moment, let us illustrate how the quantity ΔJ_B can be used as a cluster validity index.

We will consider the set of problems used in [18]. Each problem consists of n clusters, $1 \leq n \leq 5$, with each cluster containing 200 points randomly

extracted from a bivariate normal distribution whose means and covariance matrices are also randomly selected. There are 100 such data sets for each n, which makes a total of 500 different clustering problems. Figure 1 (left) shows some of the data sets for $n = 3$, $n = 4$ and $n = 5$. Using a genetic algorithm (GA) to search for the partition that minimizes ΔJ_B, one obtains the results shown in figure 1 (right).[1] Only partitions into a set of convex non-overlapping regions which are delimited by linear separators and contain at least 20 points each are allowed.

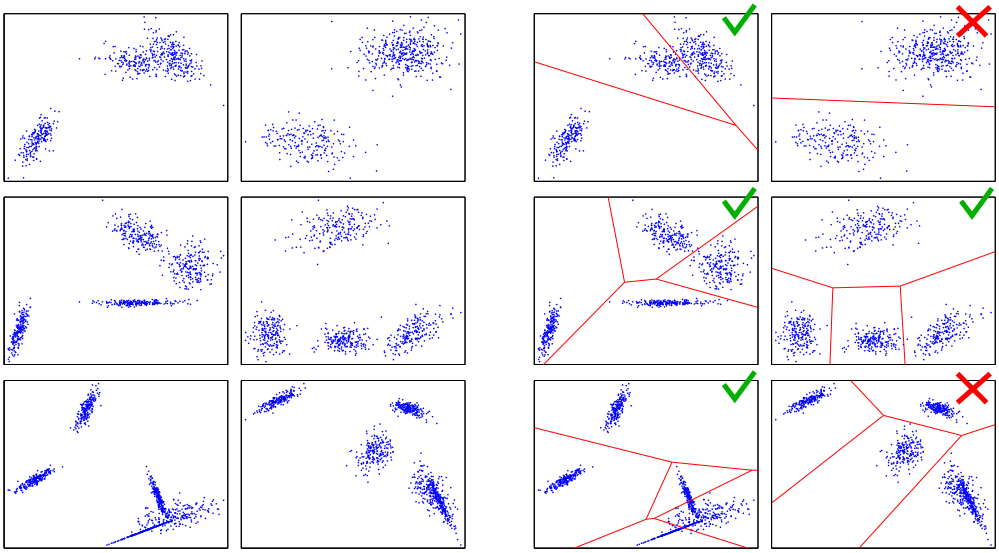

Figure 1: Left. Some examples of the problems used to test ΔJ_B for cluster validation. The number of clusters in each problem is $n = 3$ (top row), $n = 4$ (middle row), and $n = 5$ (bottom row). Right. Partitions that minimize ΔJ_B for each problem on the left. Partitions marked with a tick find the correct number of clusters. Partitions marked with a cross fail to do it. The minimum number of points in each partition region is restricted to 20.

[1]We use the PGAPack genetic algorithm library [20] with the default mutation and crossover operators. The population size is set to 500 individuals, each one representing a different partition into a set of k nonoverlapping regions. The algorithm is run for 250 iterations, and the best partition at the end is used as the solution for a particular run. We consider k values ranging from 2 to 9. We perform 20 different runs for each k, and select the solution that provides the best index value. Full details about this implementation can be found in [19].

Comparing the regions in the selected partitions to the real clusters in each problem, we observe that the ΔJ_B index is able to detect the underlying clusters with a high accuracy most of the times. To quantify this observation, we show in figure 2 (left) a plot of the number of partition regions versus the actual number of clusters. Each point in the plot is an average over the 100 problems with the same number of clusters. Note that the correct number of clusters is very close to the average estimation, being in all the cases within one standard deviation. Also note a slight overestimation on the number of clusters as n increases.

In references [18, 19] the ΔJ_B index was compared to other validity indices in the literature, showing a good performance with respect to the assessment of the number of clusters. However, as we discuss in the next section, if we drop the constraint of having at least 20 data points in each region, the previous results degrade significantly due to the presence of partition regions with a small number of points, which introduce a strong bias in the evaluation of ΔJ_B.

4. The effect of partition regions with low number of points on the estimation of ΔJ

In figure 2 (right) we plot the number of partition regions versus the actual number of clusters when there are no restrictions on the minimum number of points in the regions. As before, each point in the plot is an average over 100 problems with the same n. Now the ΔJ_B index systematically overestimates the number of clusters in about 2 units. If we draw the partitions that minimize ΔJ_B, we observe that the overall structure of the problems is detected, but some additional regions with very few points are also present (see some selected partitions in figure 3). It seems that, as far as the main clustering structure is captured, the presence of even a single region with less than 5 points is able to further decrease the value of ΔJ_B, leading to partitions that overestimate the number of clusters.

In order to understand this problem, we analyzed the size of all the regions in the selected partitions after the minimization of ΔJ_B. In figure 4 (left) we plot a histogram of this size for the partitions with the correct number of regions. We observe that there is only one main peak centered around 200, which is the number of points we generated for each cluster. As expected, the mean size of the regions coincides with the actual size of the clusters.

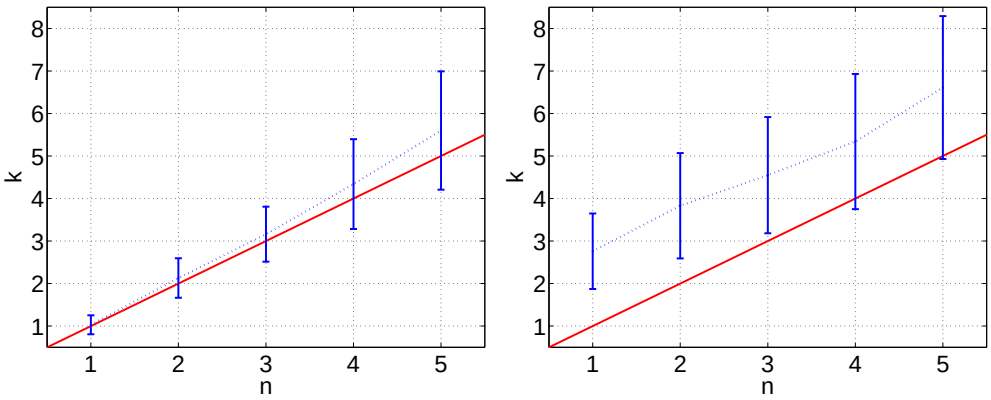

Figure 2: Average number of regions, $\overline{k}$, versus actual number of clusters, n, in the data set. Left. The minimum number of points in each partition region is restricted to 20. Right. No constraint in the number of points per region is imposed. The line $k = n$ is shown in both plots as a reference.

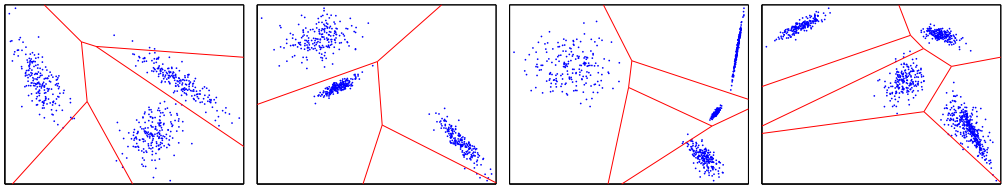

Figure 3: Some examples of the partitions found by minimization of ΔJ_B when there are no restrictions on the regions' size. In all the cases there are regions containing very few data points.

On the other hand, if we perform the same analysis for the partitions which do not correctly estimate the number of clusters, we obtain a histogram which presents two clearly separated peaks (figure 4 right). The first one is, as before, centered around 200 and corresponds to regions that correctly fit a single cluster. However there is now a second peak very close to 0, which indicates the presence of regions with only a few data points. This shows that the minimization of ΔJ_B has a strong tendency to select partitions that include regions with low number of points. This effect can be alleviated by imposing ad hoc restrictions on the minimum number of points in each region, as we showed before. A different alternative consists of correcting the bias in the estimation of ΔJ. The rest of the article follows this direction.

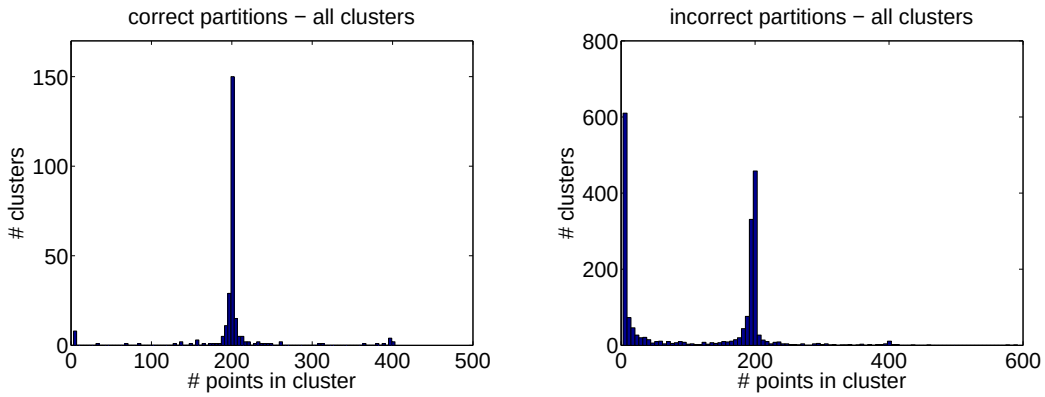

Figure 4: Histograms of regions' size for the partitions found by minimization of ΔJ_B. Left. Partitions that correctly estimate the number of clusters. Right. Partitions which find a wrong number of regions. Note that the total number of incorrect estimations is much higher than the number of correct ones.

5. Correction to ΔJ_B

In this section we show that the estimation of ΔJ given by equation 5 is biased, with the bias being stronger when there are regions in P with a small number of points. Then we show how to correct this bias. Let us call ϵ_i the error in the estimation of $\log|\Sigma_i|$:

$$\epsilon_i \equiv \log|\tilde{\Sigma}_i| - \log|\Sigma_i| \tag{7}$$

Note that $\boldsymbol{\mu}_i$ and Σ_i are the true mean and the true covariance matrix for region Ω_i; and $\tilde{\boldsymbol{\mu}}_i$ and $\tilde{\Sigma}_i$ are their sample estimations. Let us define $\mathbf{W}$ as the matrix with the eigenvectors of Σ_i (column wise), and $\mathbf{D}$ as the diagonal matrix with the corresponding eigenvalues. Then, defining $\mathbf{U} \equiv \mathbf{W}\mathbf{D}^{-1/2}$, we have $\mathbf{U}^T\Sigma_i\mathbf{U} = \mathbf{I}$. This property can be used in equation 7 to obtain:

$$\epsilon_i = \epsilon_i + 2\log|\mathbf{U}| - 2\log|\mathbf{U}| = \log|\mathbf{U}^T\tilde{\Sigma}_i\mathbf{U}| - \log|\mathbf{U}^T\Sigma_i\mathbf{U}| = \log|\mathbf{U}^T\tilde{\Sigma}_i\mathbf{U}| \tag{8}$$

where we also used the properties $|\mathbf{A}\cdot\mathbf{B}| = |\mathbf{A}|\cdot|\mathbf{B}|$ and $|\mathbf{U}| = |\mathbf{U}^T|$. On the other hand, using equation 6 the expression $\mathbf{U}^T\tilde{\Sigma}_i\mathbf{U}$ can be written as:

$$\mathbf{U}^T\tilde{\Sigma}_i\mathbf{U} = \frac{1}{N_i - 1}\sum_{\mathbf{x}\in\Omega_i}\mathbf{U}^T(\mathbf{x} - \tilde{\boldsymbol{\mu}}_i)\cdot(\mathbf{x} - \tilde{\boldsymbol{\mu}}_i)^T\mathbf{U} =$$

$$\frac{1}{N_i - 1}\sum_{\mathbf{x}\in\Omega_i}\mathbf{U}^T\left[(\mathbf{x} - \boldsymbol{\mu}_i) - (\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)\right]\cdot\left[(\mathbf{x} - \boldsymbol{\mu}_i) - (\tilde{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)\right]^T\mathbf{U} \quad (9)$$

with $\boldsymbol{\mu}_i$ being the expected value of $\mathbf{x}$ in Ω_i. The variable $\mathbf{y}$ defined as:

$$\mathbf{y} \equiv U^T(\mathbf{x} - \boldsymbol{\mu}_i) \quad (10)$$

is a random variable with expected value equal to $\mathbf{0}$ and covariance matrix equal to the identity matrix, since $\mathbf{U}^T\Sigma_i\mathbf{U} = \mathbf{I}$. Using equations 8, 9 and 10 we get:

$$\epsilon_i = \log\left|\frac{1}{N_i - 1}\sum_{\mathbf{y}\in\Omega_i'}(\mathbf{y} - \tilde{\mathbf{m}})\cdot(\mathbf{y} - \tilde{\mathbf{m}})^T\right| \quad (11)$$

where Ω_i' is the set Ω_i transformed according to equation 10, and $\tilde{\mathbf{m}}$ is the sample mean of $\mathbf{y}$. This implies the important result that the error ϵ_i follows a distribution that does not depend either on Σ_i nor $\boldsymbol{\mu}_i$. Moreover, the expected value of ϵ_i does not depend on Σ_i, being equal to the expected value of the log-determinant of the observed covariance matrix of a multidimensional Gaussian variable with zero mean and covariance matrix equal to the identity matrix.

We can use this result to show that, for any dimension and any number of points, there is a systematic bias in the estimation of $\log|\Sigma_i|$ by $\log|\tilde{\Sigma}_i|$. Let us call $\mathbf{M} \equiv \mathbf{U}^T\tilde{\Sigma}_i\mathbf{U}$. Given that the logarithm is a monotonically increasing function and that the determinant of a positive definite matrix is not greater than the product of its diagonal terms, we have:

$$E[\epsilon_i] = E[\log|\mathbf{M}|] \leq E[\log\prod_j \mathbf{M}_{jj}] = \sum_j E[\log\mathbf{M}_{jj}] \quad (12)$$

The strict version of Jensen's inequality [23] states that $E[\log x] < \log E[x]$ for a non constant random variable x. Therefore:

$$E[\epsilon_i] < \sum_j \log E[\mathbf{M}_{jj}] = 0 \quad (13)$$

since $E[\mathbf{M}] = \mathbf{I}$. This demonstrates that the estimation of ΔJ given by equation 5 is always biased towards $-\infty$. In order to study the magnitude of this bias, we performed numerical simulations to estimate the expected value, $E[\epsilon_i]$, and the standard deviation, $\sigma(\epsilon_i)$, of ϵ_i for dimension 2 and a number of points in the sample N_i ranging from 3 to 100. We generated, for each N_i, one million samples of size N_i and calculated the average and standard deviation across samples. In figures 5 (left) and 5 (right) we plot $E[\epsilon_i]$ and $\sigma(\epsilon_i)$ versus N_i. It is clear that $E[\epsilon_i] < 0$ for all N_i, tending to $-\infty$ as N_i decreases, and that $\sigma(\epsilon_i)$ tends to $+\infty$ for low N_i. To correct the bias in the estimation of the log-determinant of Σ_i we consider the following estimator:

$$T = \log|\tilde{\Sigma}_i| - E[\epsilon_i] \tag{14}$$

Note that $E[\epsilon_i]$ is a number that depends only on the dimension and the number of points in the sample N_i. It is now clear that:

$$E[T] = E[\log|\Sigma_i|] \tag{15}$$

and so the new estimator is unbiased. We can extend this analysis to show that an unbiased estimator of the negentropy increment is:

$$\Delta J_U(P) = \Delta J_B(P) + B(P) \tag{16}$$

where:

$$B(P) = \frac{1}{2}E[\epsilon_0] - \frac{1}{2}\sum_{i=1}^{k}\tilde{p}_i E[\epsilon_i] \tag{17}$$

In the next section we will consider the confidence intervals $[\Delta J_U(P) - S(P), \Delta J_U(P) + S(P)]$, where:

$$S(P) = \frac{1}{2}\sqrt{\sigma(\epsilon_0)^2 + \sum_{i=1}^{k}\tilde{p}_i^2\sigma(\epsilon_i)^2} \tag{18}$$

to determine whether the difference in ΔJ_U for two partitions is significant. For the derivation of the last expression we assumed independence between the ϵ_i. Note that, for a given dimension, the quantities $E[\epsilon_i]$ and $\sigma(\epsilon_i)$ depend only on the number of points in the region Ω_i. Since $E[\epsilon_i]$ is a monotonically

increasing function of N_i, it can be shown that the bias $B(P)$ is always negative, except for the trivial partition where all the data points belong to the same region. This demonstrates that $\Delta J_B(P)$ systematically underestimates the negentropy increment.

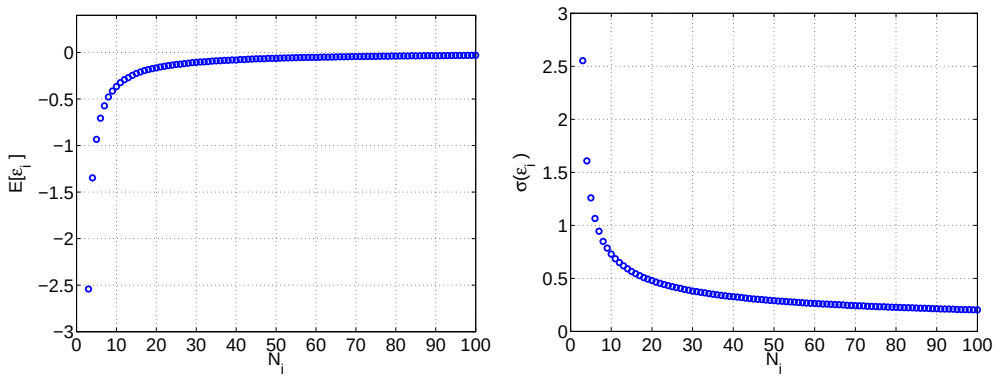

Figure 5: Numerical estimation of $E[\epsilon_i]$ (left) and $\sigma(\epsilon_i)$ (right) as a function of the sample size of region Σ_i.

6. Clustering algorithm

We test the new index ΔJ_U by using it as the fitness function of a GA that searches the partitions space, as we did in section 3 for ΔJ_B. For any clustering problem, we perform a series of runs of the GA to find the partitions[2] that minimize the ΔJ_U index for a fixed number of regions k. We consider k ranging from 2 to 9, and perform 20 different runs for each k. This provides the set of 160 partitions $\Pi = \{P_j\}, j = 1, 2, ..., 160$. On a first approach, we select as the optimal solution the partition P_{MIN} with minimum ΔJ_U across all runs:

$$P_{MIN} = P \in \Pi \ : \ \Delta J_U(P) \leq \Delta J_U(P_j) \ \ \forall \ \ j = 1, 2, ..., 160$$

As we will discuss later in the results section, this approach provides partitions that still tend to overestimate the number of clusters. We have observed that, for most of the problems where the number of clusters is

[2]To avoid numerical problems, the partitions for which some region Ω_i has $\log |\Sigma_i| \approx 0$ are discarded.

overestimated, the partition that minimizes ΔJ_U, P_{MIN}, and the optimal partition corresponding to the real clusters, P_{OPT}, have very close values of ΔJ_U. In particular there are many cases where:

$$\Delta J_U(P_{OPT}) - S(P_{OPT}) \leq \Delta J_U(P_{MIN}) + S(P_{MIN}) \tag{19}$$

In such cases, where the areas contained within one standard deviation around ΔJ_U for the two partitions overlap, we may consider that there is not enough statistical evidence to decide which partition is better, and so additional criteria must be used to select between the two partitions. An example of this situation is illustrated in figure 6.

To deal with this problem we propose a second approach:

1. Run the GA to get the set of partitions $\Pi = \{P_j\}$ and obtain, from this set, the partition P_{MIN} with minimum ΔJ_U.
2. Find the subset Π' of Π which satisfies:

$$\Pi' = \{P \in \Pi \ : \ \Delta J_U(P) - S(P) \leq \Delta J_U(P_{MIN}) + S(P_{MIN})\}$$

3. Select, from Π', the partition P with the lowest $S(P)$.

We will refer to this approach as ΔJ_{U+STD}. It guarantees that we select, for each problem, the simplest partition whose ΔJ_U is indistinguishable from $\Delta J_U(P_{MIN})$ in the sense of figure 6.

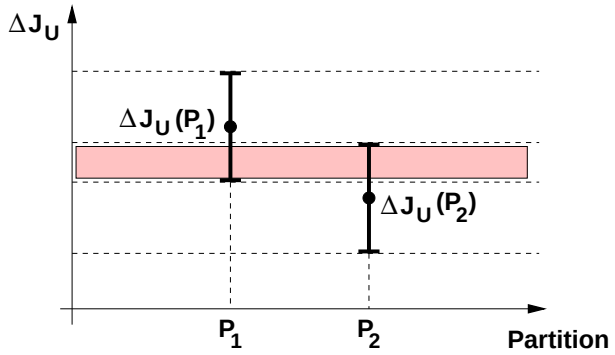

Figure 6: The partitions P_1 and P_2 are considered indistinguishable from the point of view of ΔJ_U because the areas contained within one standard deviation around $\Delta J_U(P_1)$ and $\Delta J_U(P_2)$ overlap.

7. Results

We consider two sets of problems. Firstly we use the set of randomly generated problems described in section 3. Secondly we consider two real problems from the UCI database [1]. In both cases we perform a comparison between the biased and the unbiased versions of the negentropy increment index. For a detailed comparison between the negentropy increment index and other state of the art cluster validity indices we refer the reader to [18, 19].

7.1. Gaussian clusters in 2D

Let us consider again the set of problems introduced in section 3, and let us now use the GA to find the partitions that minimize ΔJ_U instead of ΔJ_B. We impose no restrictions on the regions' size. The resulting plot of the average number of regions versus the number of clusters is shown in figure 7 (left). Although the number of clusters is still overestimated, we can observe a significant reduction in the amount of the overestimation with respect to the biased case (figure 2 right). That is, the bias correction in equation 17 is able to improve the quality of the partitions found by the GA. However these results are still not satisfactory when compared to those obtained by limiting the minimum regions' size to 20 data points (figure 2 left). It seems that there is still a strong presence of small size regions in the partitions obtained by direct minimization of ΔJ_U. On the other hand, if we use the ΔJ_{U+STD} approach to select the partitions, we obtain the results shown in figure 7 (right). Note the clear improvement with respect to all our previous results, including those where the minimum size of the partition regions was limited ad hoc. Now the average number of regions in the selected partitions estimates the real number of clusters with a higher accuracy. In the case $n = 1$ the new approach obtains the correct partition for all the problems considered.

In figure 8 we plot new histograms of the regions' size for the correct (left) and incorrect (right) partitions. Two main differences with respect to figure 4 are observed. First, the number of correct partitions is now much higher. Second, the peak close to 0 in the histogram for incorrect partitions has disappeared. This indicates that the differences between the real clustering structure of the problems and the obtained partitions is no longer due to the presence of very small regions, but to other factors related to the intrinsic difficulty of the problems (such as a high overlap or the presence of clusters that cross over each other).

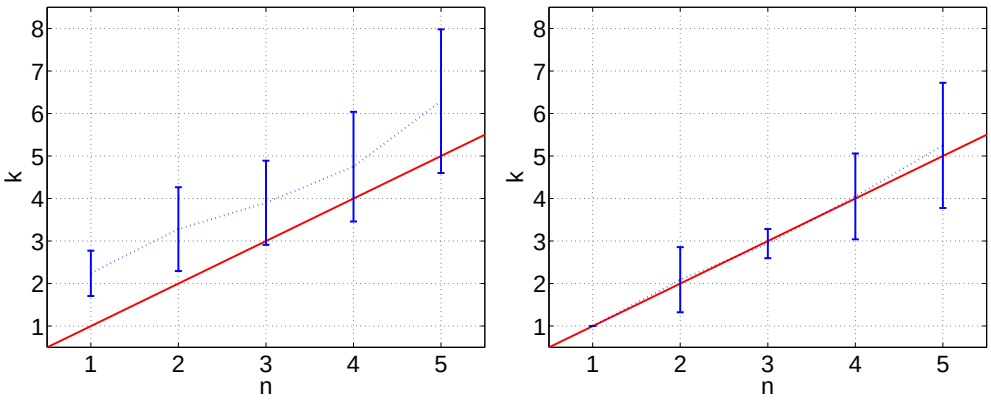

Figure 7: Average number of regions, $\overline{k}$, versus actual number of clusters, n, in the data set. Left. Results obtained by minimization of ΔJ_U. Right. Results obtained with the last approach. The line $k = n$ is shown in both plots as a reference.

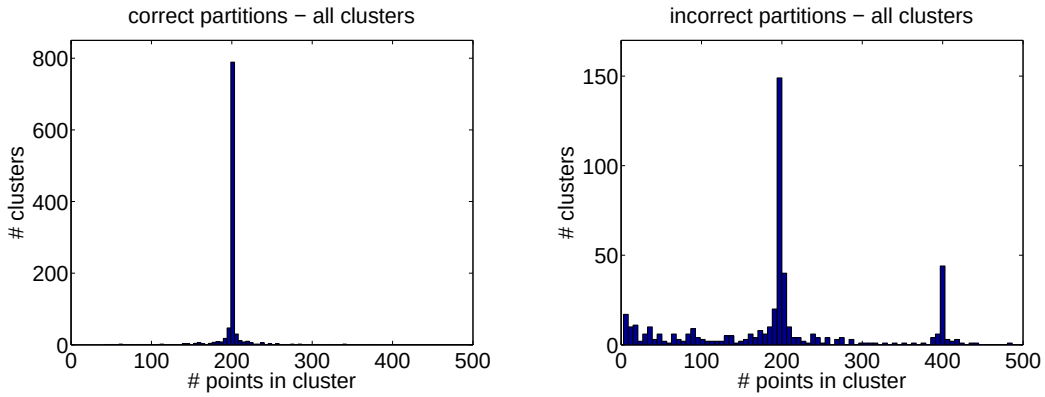

Figure 8: Histograms of regions' size for the partitions found by minimization of ΔJ_U taking into account the uncertainty $S(P)$. Left. Partitions that correctly estimate the number of clusters. Right. Partitions which find a wrong number of regions.

We performed additional analyses in order to measure the quality of the validated partitions. Previously we have focused on the correct estimation of the number of clusters. However, a good result in this estimation does not guarantee a good correspondence between the obtained partition and the real clustering structure of the problem. Thus we analyzed the discrepancy

between the predicted clusters (C_P) and the real ones (C_R), which can be measured by the entropy distance [21]:

$$D_H(C_P, C_R) = H(C_P|C_R) + H(C_R|C_P) \tag{20}$$

The entropy distance is always greater than or equal to 0, being 0 if and only if there is a one to one correspondence between C_R and C_P. In table 1 we show the average of D_H for each real number of clusters n. Each column shows the results obtained with one of the four approaches considered so far, namely validation using the biased estimator of the negentropy increment (ΔJ_B), validation using the biased estimator but limiting the minimum size of the clusters to 20 data points ($\Delta J_{B+MIN20}$), validation using the unbiased estimator (ΔJ_U), and validation using the unbiased estimator and $S(P)$ to select the simplest amongst equivalent partitions (ΔJ_{U+STD}). We observe that the last approach provides the best results.

Table 1: Average entropy distance $D_H(C_P, C_R)$ between real and predicted clusters using four different approaches.

n	ΔJ_B	$\Delta J_{B+MIN20}$	ΔJ_U	ΔJ_{U+STD}
1	0.19 ± 0.16	0.01 ± 0.08	0.13 ± 0.11	0
2	0.21 ± 0.33	0.12 ± 0.30	0.17 ± 0.30	0.10 ± 0.26
3	0.17 ± 0.21	0.13 ± 0.22	0.15 ± 0.20	0.13 ± 0.19
4	0.26 ± 0.25	0.25 ± 0.29	0.24 ± 0.24	0.24 ± 0.25
5	0.41 ± 0.24	0.39 ± 0.25	0.39 ± 0.26	0.38 ± 0.24

Finally, in order evaluate the significance of our previous results, we use the framework introduced in [9]. This framework permits to easily visualize the statistical differences among different algorithms. First, each method is ranked in each execution (rank 1 for the best method, rank 2 for the second, and so on). Then a Nemenyi test is applied to compute the statistical differences amongst the methods. Here we use the D_H values to rank each method for each one of the 500 clustering problems. The results of this test

are shown in figure 9. The average rank obtained by each method is shown in the lower axis. Methods for which the differences in average rank are not statistically significant with p-value < 0.05 are linked with a solid black line. Differences in average rank above the critical distance (CD) are considered significant. The CD is displayed at the top of the figure for reference. From this figure it can be observed that ΔJ_{U+STD} and $\Delta J_{B+MIN20}$ are the two methods with the highest performance. Although ΔJ_{U+STD} has a higher average rank, the differences between these two methods are not statistically significant. The other two methods, ΔJ_U and ΔJ_B present a much poorer performance.

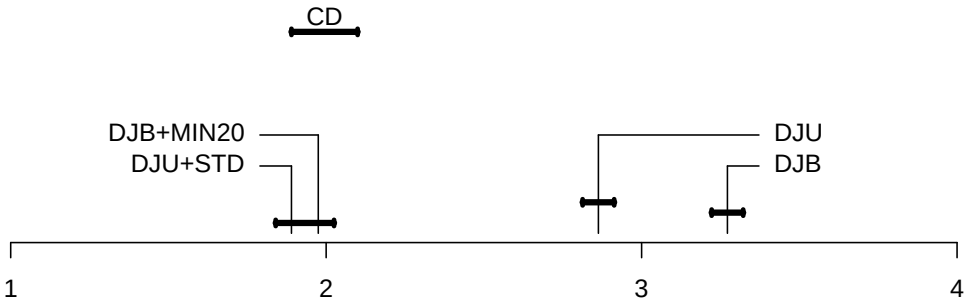

Figure 9: Average ranks of each method in the collection of 2D synthetic problems.

7.2. Real datasets

We consider two real data sets from the UCI database [1], namely the Iris data set and the Wine data set, which were also used in [19]. Although they are essentially supervised classification problems, we will use them here in an unsupervised manner.

The Iris data set consists of 150 points in a 4-dimensional attribute space. There are three classes, with 50 instances in each class. One of the classes is linearly separable from the other two. There is not agreement on whether

the number of clusters for this problem must be considered 2 or 3, so we consider here two possible solutions. The solution Iris A assumes an optimal partition into 2 clusters, one corresponding to the linearly separable class, the other containing the rest of instances. The solution Iris B assumes an optimal partition into 3 clusters, each one corresponding to one of the classes.

The Wine data set contains 178 samples characterized by 13 continuous attributes. There are 59 samples of the first class, 71 samples of the second class, and 48 samples of the third class. We have performed a PCA transformation in order to reduce the dimension to the first 6 principal components, which account for the 85.10% of the total variance.

In table 2 we compare the results for these problems obtained by the approaches ΔJ_B, ΔJ_{B+MIN} and ΔJ_{U+STD}. For the ΔJ_{B+MIN} case, the minimum number of points per partition region is set to 20 and 35 for the Iris and the Wine problems respectively. In [19], an additional requirement was necessary in order to avoid too complex solutions from the minimization of ΔJ_{B+MIN}. All the partitions with ΔJ_{B+MIN} within a 95% of the absolute minimum were considered equivalent, the simplest amongst them being selected. Here we include this requirement only for the ΔJ_{B+MIN} approach.

Our first observation is that the minimization of ΔJ_B without any additional restrictions leads to solutions exploiting the maximum allowed complexity. On the other hand, the approaches ΔJ_{B+MIN} and ΔJ_{U+STD} tend to find solutions that are close to the real clustering structure of the problems. They both obtain good partitions into 3 regions for the Wine problem. For the Iris problem, the ΔJ_{B+MIN} method finds a good partition into 3 regions that are very well related to the real classes, whilst the ΔJ_{U+STD} method finds a partition into 2 regions, one of them corresponding exactly to one of the classes (which is linearly separable from the other two). That is, the ΔJ_{B+MIN} finds the Iris B solution and ΔJ_{U+STD} finds the Iris A solution.

8. Discussion

In this article we have shown that the estimation of ΔJ presented in previous work [18, 19] is biased towards $-\infty$, this bias being stronger for partitions including regions with a small number of points. This affects the quality of the clustering partitions obtained by minimization of ΔJ which, in spite of detecting the overall structure of the problem, include additional regions with very few data points. Thus the number of clusters detected by the index tends to be overestimated. We showed that this effect can be

Table 2: Evaluation of the optimal partitions obtained with the ΔJ_B, the ΔJ_{B+MIN} and the ΔJ_{U+STD} approaches on the Iris and Wine problems, using the number of partition regions and the D_H measure.

		Num. regions			D_H		
Set	n	ΔJ_B	ΔJ_{B+MIN}	ΔJ_{U+STD}	ΔJ_B	ΔJ_{B+MIN}	ΔJ_{U+STD}
Iris A	2	9	3	2	1.23	0.46	0
Iris B	3	9	3	2	1.27	0.19	0.46
Wine	3	9	3	3	1.34	0.56	0.56

alleviated by introducing additional requirements such as limitations on the minimum number of points in each region, but these constraints need to be adjusted ad hoc depending on parameters such as the dimension of the data space.

We have formally analyzed how the estimation of the negentropy increment of a clustering partition is affected by the presence of regions with a low number of points. We found, perhaps surprisingly, that the average error in the estimation of ΔJ by ΔJ_B depends on the number of points in each region, but not on the scale or orientation of their distribution. This average error can be corrected in order to obtain an unbiased estimator ΔJ_U. Additionally, we calculated the standard deviation of the error, and used it to determine whether the difference in ΔJ_U between two different partitions is statistically significant, so that in the case of a draw the simplest partition is selected. These extensions were shown to substantially improve the quality of the clusters.

9. Acknowledgements

We thank G. Martínez-Muñoz, D. Hernández-Lobato and J.M. Hernández-Lobato for providing the software to perform the statistical analysis of the rankings. We thank the anonymous referee that provided the interesting suggestion regarding the analytical derivation of the statement $E[\epsilon_i] < 0$ for an arbitrary dimension. This work has been funded by DGUI-CAM/UAM

(project CCG10-UAM/TIC-5864).

References

[1] Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[2] Bezdek, J.C., Pal, R.N.: Some New Indexes of Cluster Validity. IEEE Trans. Systems, Man and Cybernetics B 28, 3, 301-315 (1998)

[3] Biernacki, C., Celeux, G., Govaert, G.: An Improvement of the NEC Criterion for Assessing the Number of Clusters in a Mixture Model. Pattern Recognition Letters 20, 3, 267-272 (1999)

[4] Bouguessa, M., Wang, S., Sun, H.: An Objective Approach to Cluster Validation. Pattern Recognition Letters 27, 13, 1419-1430 (2006)

[5] Bozdogan, H.: Choosing the Number of Component Clusters in the Mixture-Model Using a New Information Complexity Criterion of the Inverse-Fisher Information Matrix. In: Opitz. O., Lausen, B., Klar, R. (eds.) Data Analysis and Knowledge Organization, pp. 40-54. Springer-Verlag, Heidelberg (1993)

[6] Comon, P.: Independent Component Analysis, a New Concept? Signal Processing 36, 3, 287-314 (1994)

[7] Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley, New York (1991)

[8] Davies, D.L., Bouldin, D.W.: A Cluster Separation Measure. IEEE Trans. Pattern Analysis and Machine Intelligence 1, 4, 224-227 (1979)

[9] Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7, 1-30 (2006)

[10] Ding, Y., Harrison, R.F.: Relational Visual Cluster Validity (RVCV). Pattern Recognition Letters 28, 15, 2071-2079 (2007)

[11] Dunn, J.C.: A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters. J. Cybernetics, 3, 3, 32-57 (1973)

[12] Everitt, B., Landau, S., Leese, M.: Cluster Analysis. Hodder Arnold, London (2001)

[13] Figueiredo, M.A.T., Jain, A.K.: Unsupervised Learning of Finite Mixture Models. IEEE Trans. Pattern Analysis and Machine Intelligence 24, 3, 381-396 (2002)

[14] Geva, A.B., Steinberg, Y., Bruckmair, S., Nahum, G.: A Comparison of Cluster Validity Criteria for a Mixture of Normal Distributed Data. Pattern Recognition Letters 21, 6-7, 511-529 (2000)

[15] Gordon, A.D.: Cluster Validation. In: Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.H., Baba, Y. (eds.) Data Science, Classification and Related Methods, pp. 22-39. Springer-Verlag, New York (1998)

[16] Hathaway, R.J., Bezdek, J.C.: Visual Cluster Validity for Prototype Generator Clustering Models. Pattern Recognition Letters 24, 9-10, 1563-1569 (2003)

[17] Hyvärinen, A.: New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit. Technical Report A47, Dept. of Computer Science and Engineering and Laboratory of Computer and Information Science, Helsinki Univ. of Technology (1997)

[18] Lago-Fernández, L.F., Corbacho, F.: Using the negentropy increment to determine the number of clusters. LNCS 5517, 448-455 (2009)

[19] Lago-Fernández, L.F., Corbacho, F.: Normality-based validation for crisp clustering. Pattern Recognition 43, 782-795 (2010)

[20] Levine, D.: PGAPack Parallel Genetic Algorithm Library, http://www-fp.mcs.anl.gov/CCST/research/reports_pre1998/comp_bio/stalk/pgapack.html.

[21] D. MacKay, Information Theory, Inference and Learning Algorithms, Cambridge University Press, Cambridge, 2003.

[22] Neal, R.M.: Markov Chain Sampling Methods for Dirichlet Process Mixture Models. J. Computational and Graphical Statistics 9, 2, 249-265 (2000)

[23] Newey, W.K., McFadden, D.: Large Sample Estimation and Hypothesis Testing. In: Engle, R., McFadden, D. (eds.) Handbook of Econometrics, vol. 4. North Holland (1999)

[24] Pakhira, M.K., Bandyopadhyay, S., Maulik, U.: Validity Index for Crisp and Fuzzy Clusters. Pattern Recognition 37, 3, 487-501 (2004)

[25] Pal, N.R., Biswas, J.: Cluster Validation Using Graph Theoretic Concepts. Pattern Recognition 30, 6, 847-857 (1997)

[26] Rasmussen, C.: The Infinite Gaussian Mixture Model. In: Solla, S., Leen, T., Müller, K.-R. (eds.) Advances in Neural Information Processing Systems 12, pp. 554-560. MIT Press (2000)

[27] Richardson, S., Green, P.: On Bayesian Analysis of Mixtures with Unknown Number of Components. J. Royal Statistical Soc. B 59, 731-792 (1997)