Intelligent Data Analysis 16 (2012) 703–715
DOI 10.3233/IDA-2012-0545
IOS Press

A comparison of clustering quality indices using outliers and noise

L. Guerra[a,*], V. Robles[b], C. Bielza[a] and P. Larrañaga[a]
[a]*Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, Madrid, Spain*
[b]*Departamento de Arquitectura y Tecnología de Sistemas Informáticos, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, Madrid, Spain*

Abstract. Quality indices in clustering are used not only to assess the quality of the partitions but also to determine the number of clusters in the final result. When these indices are evaluated in a case study, real data conditions or different clustering algorithms are seldom taken into account. Here, some of the standard indices used in the literature are compared using more realistic databases that include outliers or noisy dimensions, which is more like a real problem-solving approach. Besides, three different clustering methods are used in an attempt to identify different behaviours. Also, the performance of the quality index-clustering algorithm tandem is compared to random grouping, with the aim of running an additional check. The indices are ranked, and index-based conclusions are drawn for all the scenarios.

Keywords: Clustering, internal indices, stopping rules

1. Introduction

Exploratory problems are one of the big challenges of data mining and are often tackled using clustering techniques. This type of problems are usually much more difficult to evaluate than supervised classification problems because there is no "ground truth". Because of this, the quality of each output solution is closely related to the domain problem. In spite of this, there are lots of clustering quality indices (CQIs) that try to assess the quality of the solution. To do this, indices tend to rate compact and isolated clusters highly. This partition quality is often used as a stopping rule for finding the correct number of clusters for a data set. Following this guideline, Milligan and Cooper [16] ranked 30 CQIs on an extensive battery of data sets without impurities (clear cluster structures). However, these data sets had different configurations where the number of variables and instance distribution density levels varied, that is, instances were not equally distributed in each hidden group. Hierarchical clustering was used as the approach for grouping the data. This comparison is still considered as one of the main references for clustering validation even though the work was developed 25 years ago. This is exemplified by current works dealing with the same issue [19]. These authors used the same type of data sets as [16]

*Corresponding author: L. Guerra, Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660, Boadilla del Monte, Madrid, Spain. Tel.: +34 914 524 900, ext: 1764; E-mail: l.guerra@upm.es.

but changed the evaluation approach by introducing another type of CQIs for validation. The problem is that this kind of data sets are not typical cases of real domains because of the low dimensionality and the clear separation and cohesion of the clusters. This is an utopian scenario in real problems, since there are usually irrelevant or noisy features or even no cluster structure in the data set. Throughout this document we use Milligan and Cooper's ranking as reference because of the great number of CQIs compared in that work, but there are some other references in literature that attempt to evaluate some CQIs using different clustering algorithms and kind of data, like in [22,23]. Another very recent paper comparing internal cluster validation indices is [6]. This paper attempts to obtain a standard methodology for evaluating some clustering indices using hierarchical clustering approach. They consider that it is necessary not only to get the correct number of clusters but also to obtain the partition that best fits the original data.

Regarding the CQIs and based on clustering validation literature, some authors [21] indicated that there are two groups of CQIs: internal and external. Meanwhile, other authors [7,10] discussed the existence of a third group, called relative indices. Internal validation does not require knowledge of the ground truth; the quality of a partition using internal CQIs is assessed by evaluating each individual partition based on distance or dissimilarity measures. The problem with this approach is how each partition is built, since the quality is probably measured with different criteria than were used to build the partition, which can lead to incorrect validations. On the other hand, external validation is more accurate but not realistic in clustering. In this case, the ground truth must be known and the evaluation is carried out based on this knowledge. Although there are many CQIs, some of them are equivalent [1]. The relative index concept varies depending on the above-mentioned authors. It is also interesting to note that, as indicated in [27], the CQIs suffer from biases not only with regard to the data (shape or number of clusters for instance), but also to the clustering algorithm used to obtain the partition. For this reason, and as we list in the following and detail in the next section, five different CQIs are used and, moreover, different clustering algorithms are used to partition the different kind of data.

The five internal CQIs that have been introduced in the comparison are: Silhouette [18], Calinski [3], C-index [9], DaviesBouldin (DB) [4] and Gamma [2]. Silhouette was the only CQI not compared in [16], but it is widely used in different fields, like genomics [13] or neuroscience [12] for example. The external CQI used is adjusted Rand index (ARI) [8]. This index is an improvement on the also very well-known Rand index [17].

This work focuses on using each CQI as a stopping rule for finding the real number of groups in clustering using close-to-real domains for the purpose of evaluating the proposed indices. Using the CQIs in this way is an important and common step in clustering. Thus, we compare here some of the most used CQIs in different scenarios to output some behaviour patterns that can help decision making on what index should be used depending on the problem at hand and how it is solved. The scenarios are created using different databases that aim to simulate real cases, having different percentages of outliers or noisy dimensions. Besides, data are partitioned using three different clustering approaches and one random algorithm to check if the behaviours of the indices differ in each case. Finally, and following [19], an external CQI is used as support for the validation.

The remainder of this paper is organised as follows. Section 2 presents the work method, the databases used and a brief explanation of each algorithm and index used. The experimental process is commented in Section 3. In Section 4, the experimental results are presented by the different criteria, whereas Section 5 explains the conclusions drawn from the results and some discussion.

Table 1
Summary of databases used. d is the number of dimensions, K is the number of clusters, ld are the different levels of density and n is the number of instances of each data set. Row names are the types of data sets: *clear* are non-overlapping data sets, *out5* and *out10* are data sets with 5% and 10% of outliers, respectively, finally *noi1* and *noi2* are data sets with 1 and 2 noisy dimensions, respectively

	d	K	ld	n
clear	4,6,8	2,3,4,5	1,2,3	50
out5	4,6,8,10	2,3,4,5	1,2,3	105
out10	4,6,8,10	2,3,4,5	1,2,3	110
noi1	5,7,9,11	2,3,4,5	1,2,3	100
noi2	6,8,10,12	2,3,4,5	1,2,3	100

2. Material and methods

The above five internal CQIs were compared using different data partitioning methods. Differences in data are related to impurities, such as noisy dimensions or outliers. All these details are presented in the following.

2.1. Databases used

The data sets were generated using the original cluster data generator software described in [15]. All data sets are detailed in the following, where they are divided into three groups. The first group (*clear*) is composed of data sets with strong and distinct clusters. The second group (*out5* and *out10*) has data sets generated with 5% and 10% outliers (instances that do not belong to any predefined cluster) and the last group (*noi1* and *noi2*) has data sets with 1 or 2 added random noisy dimensions (variables that do not contribute to separating the clusters).

The number of data sets in each group depends on the number of dimensions, clusters and types of density used. Thus, there are 36 data sets in the first group, resulting from combining 4 different number of clusters in the data (from 2 to 5 clusters), each with different dimensions (4, 6 or 8) and 3 different density levels designed to change the cluster sizes and the instance distributions. At the first level of density each cluster has the same number of instances; at the second level one cluster always contains 10% of instances; and at the third level one cluster contains 60% of instances. The size of each data set in this first group is 50. The second group is divided into data sets with 5% and 10% outliers. There are 48 data sets in each subgroup since a new dimensionality (10 variables) is added on top of all the combinations explained for the first group. Besides, 105 and 110 instances are used in each data set, respectively. Finally, the last two data sets have 100 instances each, but one noisy dimension is added in *noi1* and two noisy dimensions are added in *noi2*, again outputting a total of 48 data sets. Therefore, 228 data sets are used in the comparison here.

2.2. Clustering algorithms

Three clustering algorithms used in different approaches were used to partition each data set: K-means [14], hierarchical clustering [11] using Ward's method [20] and model-based clustering using Gaussian mixtures and the expectation-maximization (EM) algorithm [5]. All three algorithms are the maximum exponents of different approaches and are well-known in the clustering literature. The first two are hard clustering approaches (each instance belongs to only one cluster), whereas the model-based approach is based on soft clustering (each instance has a certain probability of belonging to each cluster). Apart from these, data were randomly partitioned as if it were another algorithm. For each data set, this random strategy was executed for 50000 iterations in an attempt to achieve statistical significance for each case.

2.3. Clustering quality indices

The study compares five internal CQIs, using one external CQI to check if clustering algorithms are able to find the correct cluster structure. All used indices were designed for being used with hard clustering algorithms, then the partitions built using model-based clustering were adapted for evaluation with the indices. There are many others CQIs that are not considered in this work, some examples are the Dunn index [24], Je(2)/Je(1) [26] or the Beale index [25].

2.3.1. Internal indices

2.3.1.1. Silhouette

The Silhouette coefficient [18], $s(i)$, is calculated for each instance i as follows:

$$s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))}, \tag{1}$$

where $a(i)$ is the average dissimilarity between instance i and all other points in the cluster to which i belongs (C for instance) and $b(i)$ is the minimum average dissimilarity to the instances of each cluster that are different to C. The average of all output values is the average Silhouette, which is the final result and is in the $[-1, 1]$ range. A high value indicates good quality clusters.

2.3.1.2. Calinski

This index [3] consists of finding well isolated clusters and is based on two measures that evaluate separation, between-cluster sum of squares (*BSS*), and cohesion, within-cluster sum of squares (*WSS*):

$$CH = \frac{BSS_K(K-1)}{WSS_K(N-K)}, \tag{2}$$

where K is the number of clusters and N is the total number of instances. The aim is to find a value of K that maximizes the index. This indicates isolated and unified clusters.

2.3.1.3. C-index

This index [9] is defined as:

$$C_{index} = \left(\frac{d_w - \min(d_w)}{\max(d_w) - \min(d_w)} \right), \tag{3}$$

where d_w is the sum of distances over all pairs of instances from the same cluster. If p is the number of pairs of instances in the same cluster, $\max(d_w)$ and $\min(d_w)$ are the sum of the p largest and smallest distances, respectively, considering all the pairs of instances. Again, this index should be minimized and is confined to the interval $[0, 1]$.

2.3.1.4. Davies-bouldin

This index [4] is calculated by averaging each pair of clusters as:

$$DB = \frac{1}{K} \sum_{i=1, i \neq j}^{K} \max \left(\frac{d_i + d_j}{d(c_i, c_j)} \right), \tag{4}$$

where K is the total number of clusters, d_i and d_j are the average distances of all instances in each cluster to their respective cluster center c_i and c_j. $d(c_i, c_j)$ is the distance between cluster centers. The target value for the DB index is small since this corresponds to compact and well-separated clusters.

2.3.1.5. Gamma

This measure is also known as Baker and Huberts index [2] and is defined as:

$$G = \left(\frac{s(+) - s(-)}{s(+) + s(-)}\right), \tag{5}$$

$s(+)$ being the number of consistent comparisons and $s(-)$ the number of inconsistent comparisons. Comparisons are made between all clusters pairwise and all between-clusters pairwise dissimilarities. A comparison is consistent if a within cluster distance is less than a between-clusters distance, otherwise it is considered as inconsistent. The target value of this index is the maximum value and it is bounded by 1.

2.3.2. External index

2.3.2.1. Adjusted rand index

The ARI [8] was created as an improvement on the Rand index [17]. The context to define these indices is: given a set of N objects, suppose $\mathcal{S}$ and $\mathcal{T}$ are two different partitions to be compared (one partition can be assumed to be the result of clustering and the other one to be the real label and we unify the name of the classes in the known partition and the clusters in the clustering results as "groups"). Then, a is the number of pairs of objects that are located in the same group in $\mathcal{S}$ and in $\mathcal{T}$, b is the number of pairs of objects in the same group in $\mathcal{S}$ but not in $\mathcal{T}$, c is the number of pairs of objects in the same group in $\mathcal{T}$ but not in $\mathcal{S}$, and d is the number of pairs of objects in different groups in both partitions $\mathcal{S}$ and $\mathcal{T}$. Then, the Rand index is defined as

$$Rand = \frac{a+d}{a+b+c+d}. \tag{6}$$

The problem of the Rand index is its value when two random partitions are compared, since it does not take a zero (minimum) value. The ARI was proposed to overcome this limitation concerning the random partitions. The ARI, like the Rand index, lies between 0 and 1, the latter being the value output when two partitions are equal. The ARI is calculated as:

$$ARI = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}. \tag{7}$$

3. Experimental process

The methodology consisted of creating the partitions, using the three clustering algorithms (K-means, hierarchical and model-based clustering) and the random grouping algorithm for all the possible number of clusters (from 2 to 5) for each data set and using the CQIs to evaluate each built partition. Thus, if 4 different algorithms and 4 cluster combinations are used for each data set, $228 * 4 * 4 = 3648$ partitions are evaluated with 5 internal and 1 external CQIs. The external CQI, ARI, is used as external validation to assess the quality of each built partition against the real partition, which is known beforehand.

For each index evaluation, the best number of clusters for each index is the maximum or the minimum value depending on the CQI. This choice will be correct if the chosen number of clusters matches the real number of clusters known beforehand. Otherwise, the choice will be classed as wrong irrespective of the distance to the real number of clusters. Besides, evaluating the external CQI, it is possible to find

Table 2
Number and percentage of correct decisions for each CQI, clustering algorithm and number of clusters (from 2 to 5) in 228 data sets: 96 with outliers, 96 with noise and 36 with clear cluster structures

		2	3	4	5	total	%
Silhouette	K-means	31	31	29	33	124	54.386
	Hierarchical	38	39	35	32	144	63.158
	EM based	39	33	31	33	136	59.649
Calinski	K-means	50	40	28	42	160	70.175
	Hierarchical	51	39	29	40	159	69.737
	EM based	50	37	33	36	156	68.421
C-index	K-means	22	14	13	31	80	35.088
	Hierarchical	9	8	7	35	59	25.877
	EM based	6	4	7	39	56	24.561
DB	K-means	27	31	28	31	117	51.316
	Hierarchical	30	35	31	28	124	54.386
	EM based	41	32	25	21	119	52.193
Gamma	K-means	21	32	29	36	118	51.754
	Hierarchical	23	36	30	36	125	54.825
	EM based	22	30	27	41	120	52.632
ARI	K-means	55	55	52	46	208	91.228
	Hierarchical	57	55	52	47	211	92.544
	EM based	55	55	54	49	213	93.421

out if the clustering algorithms were able to find the real clusters structure for each data set in spite of correct or incorrect choices by the internal CQIs of the number of clusters.

All these points are evaluated on the above-mentioned data types, outputting results for data with non-overlapping (*clear*) clusters. Data with outliers and data with noisy dimensions are then added to compare the behaviour of the CQIs with these data types.

4. Results

The first results are presented in Table 2. Table 2 shows the number of correct decisions on the number of clusters output by each CQI and each clustering algorithm in all the databases used (228 data sets). Clearly, the number of correct decisions output by the ARI is very high. This means that the clustering algorithms were able to find the correct cluster structure in many situations, especially when they had to find 2 or 3 clusters. In the case of internal indices, Calinski achieved the best results with around 70% of correctly identified clusters. It was followed by, Silhouette, DB and Gamma, which all achieved very similar results. C-index was the clear loser in this first comparison. It was interesting to note however that this index was at least as competitive as Silhouette, DB or Gamma at finding 5 clusters.

Regarding the clustering algorithms, K-means behaved better when used with Calinski and C-index, whereas hierarchical clustering outperformed the other algorithms when used with Silhouette, DB or Gamma. Interestingly, EM was the best algorithm only when used with ARI, which was the most precise index.

4.1. Index behaviour in data sets with outliers

The results using the 96 data sets with 5% and 10% of outliers are shown in Table 3. They were different from the general results shown in Table 2. In this case, the biggest differences depended

Table 3
Number and percentage of correct decisions for each CQI, clustering algorithms and number of clusters in 96 data sets with 5% and 10% of outliers

		2	3	4	5	total	%
Silhouette	K-means	3	7	8	16	34	35.417
	Hierarchical	9	14	14	16	53	55.208
	EM based	6	7	10	18	41	42.708
Calinski	K-means	18	19	11	21	69	71.875
	Hierarchical	18	18	13	20	69	71.875
	EM based	17	14	13	17	61	63.542
C-index	K-means	12	11	9	17	49	51.042
	Hierarchical	3	6	4	21	34	35.417
	EM based	0	1	2	22	25	26.042
DB	K-means	2	6	9	16	33	34.375
	Hierarchical	7	9	13	13	42	43.750
	EM based	18	6	9	5	38	39.583
Gamma	K-means	2	5	5	13	25	26.042
	Hierarchical	3	6	6	15	30	31.250
	EM based	1	2	5	18	26	27.083
ARI	K-means	22	22	22	20	86	89.583
	Hierarchical	24	22	23	23	92	95.833
	EM based	22	22	24	24	92	95.833

on the clustering algorithm used. For example, there was a 20% difference in the number of correct decisions using K-means and hierarchical clustering in Silhouette. This also applies to C-index, where EM returned around 26% and K-means around 51% correct decisions. In any case, the internal CQI with the highest percentage of correct decisions was again Calinski, but, taking into account all three clustering algorithms, Gamma had a worse mean than C-index. This algorithm achieved a better result than before thanks mainly to the improvement in the K-means output.

The behaviour of clustering algorithms with each CQI was very similar to before, and there were no significant differences. A minor difference was that K-means used with C-index performed much better than the other clustering algorithms. Proportionately, the other differences among the three clustering algorithms were unchanged.

With the appearance of outliers, the results for the internal CQIs were very poor. Again, ARI output high outcomes, which means that the clustering algorithms found the cluster structures. However, the internal CQIs were not able to find the structures used as stopping rules to determine the number of clusters.

4.2. Index behaviour in data sets with noise

The results using the 96 data sets with 1 and 2 dimensions of noise are shown in Table 4. In general, the results of the internal CQIs improved in data sets with these characteristics compared with outliers. This applies for Silhouette, DB and Gamma. The difference in Calinski was more balanced, whereas C-index was the big loser in data sets with noise. The number of correct decisions on the number of clusters using C-index was very low, mainly when the number of clusters was not 5. Because of these results, C-index was the lowest ranked internal CQI in this comparison. On the other hand, Gamma, which was the index with the poorest results in data sets with outliers, achieved the best results in data sets with noise.

Table 4
Number and percentage of correct decisions for each CQI, clustering algorithms and number of clusters in 96 data sets with 1 and 2 noisy dimensions

		2	3	4	5	total	%
Silhouette	K-means	20	19	13	11	63	65.625
	Hierarchical	21	19	13	10	63	65.625
	EM based	24	20	14	9	67	69.792
Calinski	K-means	24	15	12	13	64	66.667
	Hierarchical	24	15	11	12	62	64.583
	EM based	24	16	15	11	66	68.750
C-index	K-means	5	1	3	12	21	21.875
	Hierarchical	3	0	2	12	17	17.708
	EM based	3	1	3	14	21	21.875
DB	K-means	18	19	12	9	58	60.417
	Hierarchical	16	19	11	9	55	57.292
	EM based	16	19	10	9	54	56.250
Gamma	K-means	14	22	15	16	67	69.792
	Hierarchical	15	24	15	14	68	70.833
	EM based	15	22	14	16	67	69.792
ARI	K-means	24	24	21	18	87	90.625
	Hierarchical	24	24	20	16	84	87.500
	EM based	24	24	22	17	87	90.625

Another interesting feature of these results was the improvement of EM, which outperformed K-means and hierarchical clustering when used with Silhouette or Calinski and was at least as competitive as the others when used with C-index or Gamma.

4.3. CQI evolution depending on the data sets

Another important aspect to be examined in this work is how each CQI evolves when data change from "clean" clusters to data with outliers or noisy dimensions. This could lead to conclusions about how these new data characteristics affect the behaviour of a CQI and determine when it a particular combination of CQI and clustering algorithm is better.

The complete evolution is shown in Fig. 1. One of the most interesting findings was that C-index performed worse with the *clear* data sets than with outliers. However, when Calinski was used to find the correct number of clusters in data sets with 5% outliers, the outcomes were at least as competitive as in *clear* data sets. One important point for examination here was how the introduction of more outliers or noisy dimensions affected the behaviour of each index. Silhouette performed worse in data sets with outliers than in data sets with noise, but the introduction of the second noisy dimension had a bigger effect than the switch 5% to 10% outliers. This situation was even more marked using Calinski, since the performance decreased substantially compared with the other data sets when the second noisy dimension was introduced. DB and Gamma behaved similarly: results for data sets with outliers were very poor, whereas values for one noisy dimension data sets were competitive compared with *clear* data sets. When the second noisy dimension was introduced, the performance decreased. The exception was the C-index discussed above. Performance with this index was generally very low, but the results with K-means and hierarchical clustering were better when outliers were introduced. This was exception among internal CQIs, but not compared with ARI, since the external CQI performed better in data sets with outliers than in data sets with noise.

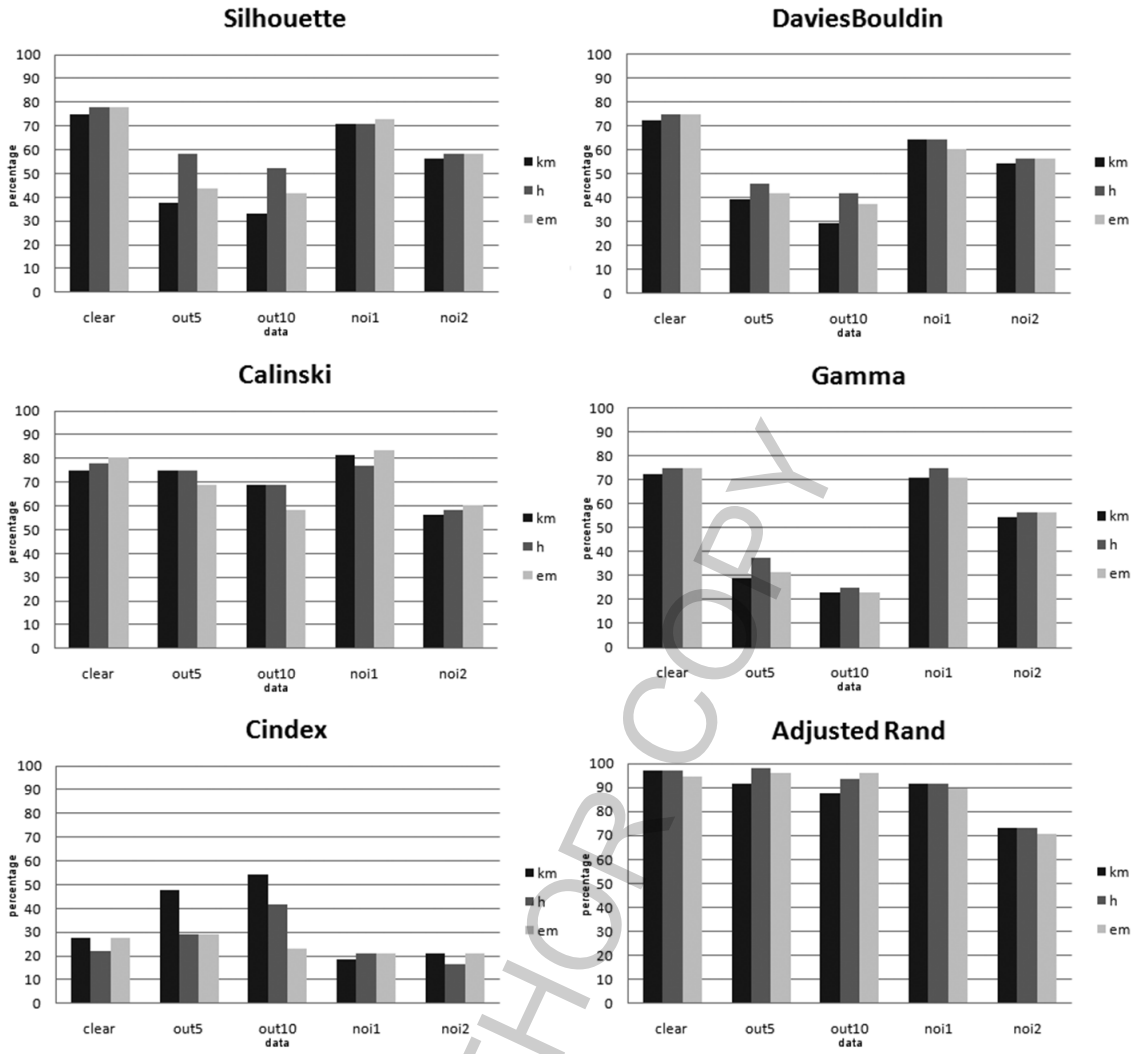

Fig. 1. Evolution of the percentage of correct decisions in the number of clusters of each CQI depending on the type of data sets and the clustering algorithms.

Regarding the algorithms, EM and hierarchical clustering achieved the best results compared with K-means for clear data sets. For data sets with outliers, hierarchical clustering achieved results that were at least as good as the other clustering algorithms with all the internal CQIs, except for C-index, where K-means was the winner. In data sets with noisy dimensions, there was not a very clear pattern of clustering algorithm behaviour. The top cluster algorithms changed depending on the CQI and whether there was 1 or 2 noisy dimensions.

Besides the number of correct decisions on the number of clusters in a data set, another factor for evaluation was how the value of each index changed depending on the characteristics of the data. The mean values of each CQI for each data situation are shown in Fig. 2. Remember that the aim of Silhouette, Calinski, Gamma and ARI is to maximize the value, whereas the lower the value of C-index and DB the better. In general, the addition of noisy dimensions affected the values of all the indices more, and

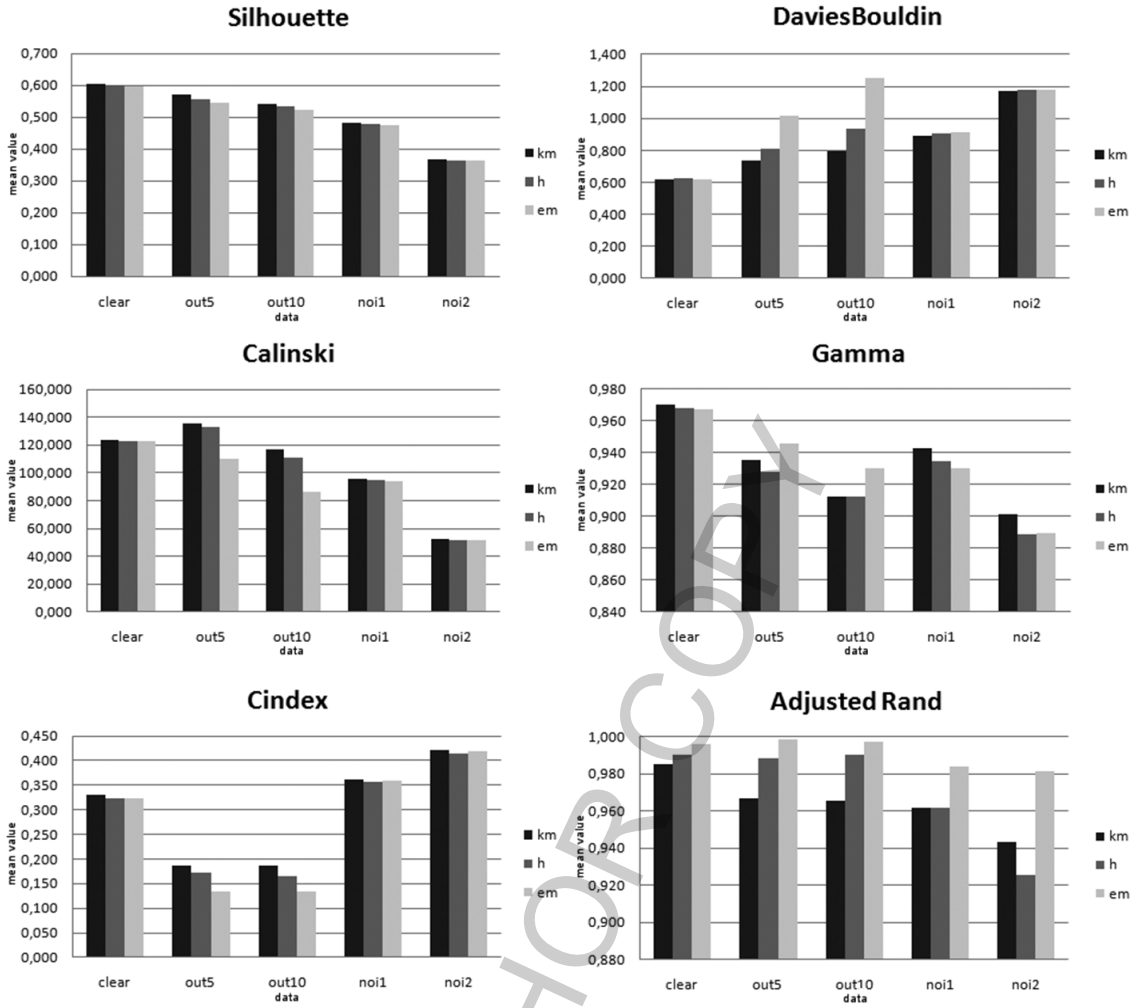

Fig. 2. Mean values of each CQI for each type of data set. Note that the scale is different depending on the index.

performance was worse. C-index was again an exception, because the values for the *clear* data sets were worse than for data sets with outliers. Note also that, when outliers were introduced in C-index previously, this index returned a better percentage of correct decisions, here again, when index value is observed, the introduction of outliers improved the behaviour of the C-index value. As regards how the introduction of more outliers or more noisy dimensions affected the output values, the performance of Calinski and Gamma decreased considerably when the second noisy dimension was introduced. In general, the addition of 5% of outliers affected the performance of all indices except for C-index and ARI.

4.4. Random groups and clustering algorithms

After running the studies using different clustering algorithms to partition the data, a different approach was introduced. Data was randomly partitioned to compare the behaviour of the CQIs with the returned values previously when clustering algorithms were used. There are 50000 random partitions for each

Table 5
Percentage of cases in which the different internal CQIs output better evaluation scores for random partitioning than for the clusterin algorithms

		clear	out5	out10	noi1	noi2
Silhouette	K-means	0.21	0.39	0.45	0.49	0.66
	Hierarchical	0.21	0.41	0.47	0.49	0.66
	EM based	0.21	0.41	0.44	0.49	0.66
Calinski	K-means	8.23	6.78	5.77	10.92	0.61
	Hierarchical	7.32	6.28	4.91	11.13	0.66
	EM based	8.91	6.32	6.48	9.82	0.64
C-index	K-means	8.58	13.17	11.20	9.17	4.58
	Hierarchical	6.42	5.25	2.66	7.61	4.27
	EM based	7.95	0.59	3.21	7.72	5.41
DB	K-means	0.36	1.21	1.80	3.12	4.10
	Hierarchical	0.35	2.65	4.69	3.37	4.16
	EM based	0.19	5.41	8.85	3.30	4.31
Gamma	K-means	0.15	0.23	0.22	0.25	0.24
	Hierarchical	0.15	0.23	0.22	0.25	0.24
	EM based	0.15	0.23	0.22	0.25	0.24

data set, outputting the same number of quality assessments for each CQI analyzed in the study. As this was a random approach, we consider the specified number of repetitions in an attempt to achieve statistical significance.

In this case, we are not interested in how similar the random partitions are to the original groups, and the external CQI is not used to assess this aspect. The aim of evaluating random partitions with the internal CQIs is to compare these assessments with evaluations using the clustering algorithms and output the percentage of random executions that scored better values than clustering algorithm validations.

The results are shown in Table 5. Table 5 shows that Silhouette and Gamma were the two more logical indices. Logical means indices whose partitioning evaluation results with clustering algorithms were not usually outperformed (less than 0.66% of times) by random partitioning. On the other hand, random partitions assessed with Calinski, C-index and DB indices scored different percentages depending on the data type and the clustering algorithm. In the "Calinski-data type" tandem, when *noi2* data sets are partitioned, clustering algorithms are beaten only by a maximum of 0.66% of executions. In all the other data sets, this percentage is significantly greater, ranging from 4.91% of random partitions, which beat the score with hierarchical clustering in *out10* data sets, to 11.13% using the same algorithm to partition *noi1* data set. In the case of C-index, when this index evaluated partitions output with K-means, the results were beaten by random partitioning from a 4.58% of cases in *noi2* data sets to 13.17% in *out5* data sets. When *out5* data sets are partitioned with the EM-based clustering algorithm, the percentage of cases in which random partitioning scored a better value for C-index is 0.59%; in all other cases using EM or hierarchical clustering, this value increased up to a maximum of 7.95%. Random partitioning evaluated with the DB index outperformed the clustering algorithms in fewer than 1% of the cases for *clear* data sets, but again this value increased up to 8.85% for *out10* data sets with EM-based clustering.

Another interesting result was the clear tendency of each internal CQI to choose an "extreme" number of clusters as correct. In the case of Silhouette and C-index, this number of clusters was 2 (which was the minimum number of clusters used in this study). On the other hand, Calinski and DB tended to choose 5 clusters, which was the maximum number of clusters. Gamma was the only index that was not so biased to a set number of clusters.

Table 6
On a scale of 1–5, where 5 is the best, the different studied parameters are summarized for each internal CQI. Note that to determine each value 1–5, numeric results were placed in 5 bins. In case of random stability and bias to min/max number of clusters, the value was approximated depending on the conclusions

	general behaviour	outliers effect	noise effect	random stability	bias to min/max # of clusters
Silhouette	3	3	4	4	2
Calinski	4	4	4	2	2
C-index	2	3	1	2	2
DaviesBouldin	3	2	3	3	2
Gamma	3	2	4	4	4

5. Conclusions

The work of Milligan and Cooper [16] established an interesting ranking presenting a lot of internal CQIs. A conclusion of that work was that the data set was an important factor influencing the good or bad behaviour of internal CQIs. This is a recurrent scenario, and it is very hard to find a dataset independent index. In spite of this, it is important to know how the most important indices behave in more complex databases to be able to choose one or another depending on the characteristics of the problem.

This work throws light on the behaviour of some of the best known index is used to cluster data sets with outliers or noise. Some conclusions that should be taken wisely were that C-index was the worse index presented here due to its generally low precision, unchanged even when noisy dimensions were added to the database. Besides, this index does not behave as expected, because it achieved better results in data sets with outliers than in data sets without outliers or noisy dimensions. In Milligan and Cooper's ranking, C-index was placed third, but this index does not appear to be so interesting if the data set contains impurities. On the other hand, Calinski achieved the best mean of correct decisions. Note, however, how the performance dropped considerably using this index when the second noisy dimension was added to the database. It dropped again when the number of dimensions was incremented.

Another important issue was to find a stable index. In this respect, Gamma was not outperformed by random groups, was not biased to a set number of clusters and was not so much affected by the introduction of the second noisy dimension did as the other indices. Silhouette's performance was not bad when noisy dimensions were added, but it was affected by outliers and also by the increased number of dimensions. A similar thing applies to the DB index, but its quality is generally slightly lower. Considering the behaviour of each index used with different clustering algorithms, it is difficult to find clear patterns indicating the best algorithm-index combinations. In general, hierarchical clustering returned more promising results than K-means or the EM algorithm, but this should not be seen as a categorical conclusion because the differences depended on the characteristics of each data set. Based on these conclusions, if these indices were to be ranked, Gamma would placed first, followed by Silhouette and Calinski, with DB one step below and C-index at the bottom of the ranking. Finally, in an attempt to clarify the conclusions drawn from the study, Table 6 summarizes the studied parameters for each internal CQI.

All these conclusions are obtained using data with outliers and with noisy dimensions as indicated. A possible future line of this work is to include other kinds of data, like data with missing values for example. Also following this line, other clustering algorithms can be used to partition the data. Assessing the quality of clustering and finding the correct number of clusters in a partition is a very open field, but this work provides new guidelines on for using a specific index depending on the characteristics of the data set.

Acknowledgments

This research is partially supported by the Spanish Ministry of Science and Innovation project TIN2010-20900-C04-04, the Cajal Blue Brain project and also Consolider Ingenio 2010-CSD2007-00018.

References

[1] A.N. Albatineh, M. Niewiadomska-Bugaj and D. Mihalko, On similarity indices and correction for chance agreement, *Journal of Classification* **23** (2006), 301–313.
[2] F.B. Baker and L.J. Hubert, Measuring the power of hierarchical cluster analysis, *Journal of the American Statistical Association* **70** (1975), 31–38.
[3] T. Calinski and J. Harabasz, A dendrite method for cluster analysis, *Communications in Statistics* **3**(1) (1974), 1–27.
[4] D.L. Davies and D.W. Bouldin, A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1** (1979), 224–227.
[5] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society* **39**(1) (1977), 1–38.
[6] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, J.M. Pérez and J.I. Martín, Towards a standard methodology to evaluate internal cluster validity indices, *Pattern Recognition Letters* **32**(3) (2011), 505–515.
[7] M. Halkidi, Y. Batistakis and M. Vazirgiannis, On clustering validation techniques, *Journal of Intelligent Information Systems* **17** (2001), 107–145.
[8] L. Hubert and P. Arabie, Comparing partitions, *Journal of Classification* **2**(1) (1985), 193–218.
[9] L.J. Hubert and J.R. Levin, A general statistical framework for assessing categorical clustering in free recall, *Psychological Bulletin* **83** (1976), 1072–1080.
[10] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Prentice Hall College Division, 1988.
[11] S.C. Johnson, Hierarchical clustering schemes, *Psychometrika* **2** (1967), 241–254.
[12] A. Karagiannis, T. Gallopin, C. David, D. Battaglia, H. Geoffroy, J. Rossier, E.M.C. Hillman, J.F. Staiger and B. Cauli, Classification of NPY-Expressing neocortical interneurons, *Journal of Neuroscience* **29**(11) (2009), 3642–3659.
[13] L. Lovmar, A. Ahlford, M. Jonsson and A.C. Syvanen, Silhouette scores for assessment of SNP genotype clusters, *BMC Genomics* **6**(1) (2005), 1–35.
[14] J. MacQueen, Some methods for classification and analysis of multivariate observations, In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 1967, pp. 281–297.
[15] G.W. Milligan, An algorithm for generating artificial test clusters, *Psychometrika* **50**(1) (1985), 123–127.
[16] G.W. Milligan and M. Cooper, An examination of procedures for determining the number of clusters in a data set, *Psychometrika* **50**(2) (1985), 159–179.
[17] W.M. Rand, Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association* **66**(336) (1971), 846–850.
[18] P. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* **20**(1) (1987), 53–65.
[19] L. Vendramin, R.J.G.B. Campello and E.R. Hruschka, On the comparison of relative clustering validity criteria, In *Proceeding of the Ninth SIAM International Conference on Data Mining*, 2009, pp. 733–744.
[20] J.H. Ward, Hierarchical grouping to optimize an objective function, *Journal of American Statistical Association* **58**(301) (1963), 236–244.
[21] K.Y. Yeung, D.R. Hayunor and W.L. Ruzzo, Validating clustering for gene expression data, *Bioinformatics* **17**(4) (2001), 309–318.
[22] U. Maulik and S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(12) (2002), 1650–1654.
[23] E. Dimitriadou, S. Dolnicar and A. Weingessel, An examination of indexes for determining the number of clusters in binary data sets, *Psychometrika* **67**(3) (2002), 137–160.
[24] J.C. Dunn, Well separated clusters and optimal fuzzy-partitions, *Journal of Cybernetics* **4** (1974), 95–104.
[25] E.M.L. Beale, Cluster analysis, *London: Scientific Control Systems*, 1969.
[26] R.O. Duda and P.E. Hart, Pattern classification and scene analysis, New York, Wiley, 1973.
[27] J. Handl, J. Knowles and D.B. Kell, Computational cluster validation in post-genomic data analysis, *Bioinformatics* **21**(15) (2005), 3201–3212.