Towards a Standard Methodology to Evaluate Internal Cluster Validity Indices

Ibai Gurrutxaga[a], Javier Muguerza[a], Olatz Arbelaitz[a], Jesús M. Pérez[a], José I. Martín[a]

[a] *Department of Computer Architecture and Technology, University of the Basque Country, 20018 Donostia, Spain*

Abstract

The evaluation and comparison of internal cluster validity indices is a critical problem in the clustering area. The methodology used in most of the evaluations assumes that the clustering algorithms work correctly. We propose an alternative methodology that does not make this often false assumption. We compared 7 internal cluster validity indices with both methodologies and concluded that the results obtained with the proposed methodology are more representative of the actual capabilities of the compared indices.

Key words: Clustering, Cluster Validity Index

1. Introduction

Clustering is an unsupervised pattern classification method that partitions the input space into groups or clusters. The goal of a clustering algorithm is to perform a partition where objects within a group are similar and objects in different groups are dissimilar. Therefore, the purpose of clustering is to identify natural structures in a dataset (Jain and Dubes, 1988; Halkidi et al., 2001; Mirkin, 2005; Sneath and Sokal, 1973) and it is widely used in many fields such as psychology (Holzinger and Harman, 1941), biology (Sneath and Sokal, 1973), pattern recognition (Mirkin, 2005), image processing (Chou et al., 2004) and computer security (Barbará and Jajodia, 2002).

Once a clustering algorithm has processed a dataset and a partition of the input data is obtained, a relevant question arises: How well does the partition fit the data? This question is important for two reasons. First, an optimal clustering algorithm does not exist. That is to say, different algorithms, or different configurations of the same algorithm, produce different partitions and none of them have proved to be the best in all situations (Pal and Biswas,

Email addresses: i.gurrutxaga@ehu.es (Ibai Gurrutxaga), j.muguerza@ehu.es (Javier Muguerza), olatz.arbelaitz@ehu.es (Olatz Arbelaitz), txus.perez@ehu.es (Jesús M. Pérez), j.martin@ehu.es (José I. Martín)

1997). Thus, in an effective clustering process we should compute different partitions and select the one that best fits the data. Secondly, many clustering algorithms are not able to determine the number of natural clusters in the data, and therefore they must initially be supplied with this information. Since this information is rarely previously known the usual approach is to run the algorithm with different values and select the partition that best fits the data. The process of estimating how well a partition fits the structure underlying the data is known as cluster validation (Halkidi et al., 2001).

For validating data partitions we must examine the clusters determined by the evaluated partition and measure the compactness of the clusters and their separation. Many authors have proposed different indices, called internal cluster validity indices (Halkidi et al., 2001; Kim and Ramakrishna, 2005; Maulik and Bandyopadhyay, 2002), to perform this validation. Unfortunately, no internal CVI has proved to be efficient in all conditions. In the rest of the paper we will refer to them as cluster validity indices or CVIs.

Many authors have compared the accuracy of cluster validity indices proposed in the literature (Dimitriadou et al., 2002; Maulik and Bandyopadhyay, 2002; Milligan and Cooper, 1985). In addition, most new CVI proposals have been compared to well known indices (Chou et al., 2004; Hardy, 1996; Kim and Ramakrishna, 2005; Kothari and Pitts, 1999; Pal and Biswas, 1997). Therefore, it is clear that a method for comparing different CVIs, which we call an evaluation of cluster validity indices, is necessary. Although little theoretical work has been done in this context, most CVI evaluation work has followed the same methodology (Chou et al., 2004; Devillez et al., 2002; Günter and Bunke, 2003; Kim and Ramakrishna, 2005; Maulik and Bandyopadhyay, 2002; Milligan and Cooper, 1985; Pal and Biswas, 1997). We call this methodology the classical methodology.

We hypothesize that the classical methodology is based on an incorrect assumption. Moreover, we have developed an alternative methodology to evaluate cluster validity indices, which overcomes this problem. Results obtained from the evaluation of 7 well-known CVIs on 10 real and synthetic datasets supported the suitability of the proposed methodology.

The classical methodology and its fundamental assumption are briefly described in the next two sections. We review some previous work that evaluates CVIs in Section 4. In Section 5 we present the alternative CVI evaluation methodology. Sections 6 and 7 are devoted describing an empirical comparison between the classical CVI evaluation methodology and the proposed alternative. Finally, conclusions are drawn in Section 8.

2. Classical methodology to evaluate CVIs

In order to evaluate a group of CVIs we need a set of datasets and a clustering algorithm. We need to know the exact number of clusters for each dataset. Therefore, the datasets are usually synthetic and low-dimensional, commonly 2-dimensional, so we can visually check the correct number of clusters. The clustering algorithm must allow an input parameter that sets the number of clusters

the generated partition will have, also known as the k parameter due to the well known k-means algorithm. The agglomerative hierarchical algorithm (Jain and Dubes, 1988) and the k-means algorithm (Mirkin, 2005) are widely used for this purpose (Kim and Ramakrishna, 2005; Maulik and Bandyopadhyay, 2002; Milligan and Cooper, 1985; Pal and Biswas, 1997).

The algorithm is run over the dataset with a set of m different values for the k parameter, $K = \{k_1, k_2 \ldots, k_m\}$. In this way, a set of m partitions is obtained, $S = \{P_1, P_2, \ldots, P_m\}$, but just one of them has partitioned the data with the correct number of clusters. We refer to this particular partition as the P^{nc} partition.

$$P^{nc} = P_i | \, \mathrm{nc}(P^*) = \mathrm{nc}(P_i), P_i \in S$$

where P^* is the correct partition of the analysed dataset, and $\mathrm{nc}(P)$ the number of clusters of a partition P.

The CVI is computed for all the partitions in S and the partition obtaining the best value for the evaluated CVI will serve to predict the actual number of clusters. For the sake of simplicity let us assume that the index used assigns greater values to "better" partitions. If the function $\mathrm{I}(P)$ computes the value obtained by the evaluated index over the partition P, we say that the cluster validity index proposes partition P^I as the best partition in S.

$$P^I = \arg\max_{P_i \in S}(\mathrm{I}(P_i))$$

We say that the index has predicted that the dataset contains $\mathrm{nc}(P^I)$ clusters and consider that it has made a successful guess if $\mathrm{nc}(P^I) = \mathrm{nc}(P^*)$. In the evaluation the more times a CVI guesses the number of clusters of the different datasets the better it is considered to be.

This type of work is data dependent since different indices behave in a different manner on different datasets. Despite these limitations, very interesting conclusions can be drawn from this type of study (Milligan and Cooper, 1985).

3. The algorithm correctness assumption

The CVI evaluation process described above works under a fundamental assumption: the clustering algorithm works "correctly"; or to be more precise, of the m partitions that the algorithm has determined for a particular dataset, the P^{nc} partition is the one that best fits the data. If this assumption does not hold —there is a partition $P_i \in S$ that fits the data better than the P^{nc} partition— it is not fair to ask the CVI to guess the actual number of clusters. Since the CVI is designed to measure the correctness of a partition, in these anomalous situations the prediction of an incorrect number of clusters is not just a forgivable error but a desirable behaviour.

This fundamental assumption is obviously true when the algorithm is able to perfectly find the structure underlying the data ($P^* \equiv P^{nc}$), and it should be true when the algorithm shows close to optimal behaviour. However, it is

difficult for the assumption to hold in complex environments, such as noisy datasets, overlapped clusters, non-convex clusters, etc.

In this section we show some examples to illustrate that the above mentioned assumption does not hold in many situations; classical clustering algorithms frequently fail even on datasets with compact and well separated clusters.

In Figure 1a we present a noisy 2-dimensional dataset. There are 4 compact and well separated clusters and some noise added under a random uniform distribution. We used the single-linkage agglomerative hierarchical algorithm to partition the data and Figure 1b shows the P^{nc} partition obtained; that is, the partition corresponding to the 4-cluster solution proposed by the algorithm. We found that 3 clusters were joined together while two noise points were proposed as singleton clusters. The singleton clusters are circled in the figure. The same algorithm is able to perfectly partition the non-noise data if k is set to 10. We consider that a CVI should favour this 10-cluster partition and not the incorrect 4-cluster one.

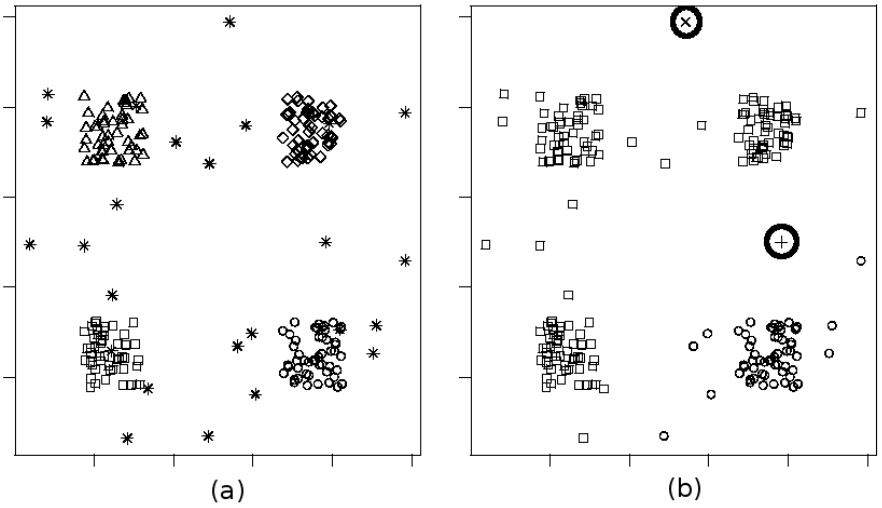

Figure 1: a) *Noisy* dataset. 4 compact clusters and some noise points. b) The 4-cluster partition proposed by the single-linkage hierarchical clustering algorithm.

In Figure 2a we show a 2-dimensional dataset with just 2 clusters. In this case no noise is present. We run the k-means algorithm over this data with $k = 2$ and we obtained the partition shown in Figure 2b . Once again, the partition was not correct. In this case the reason is that the k-means has hard problems in correctly partitioning non-convex clusters. Since the k-means algorithm has a random component (the centroid initialization) we ran the algorithm 10 times with $k = 2$ and 10 times with $k = 3$. 9 of the partitions with 3 clusters fit the data better than any of the partitions with 2 clusters. Once again, we consider that a CVI should favour these 3-cluster partitions.

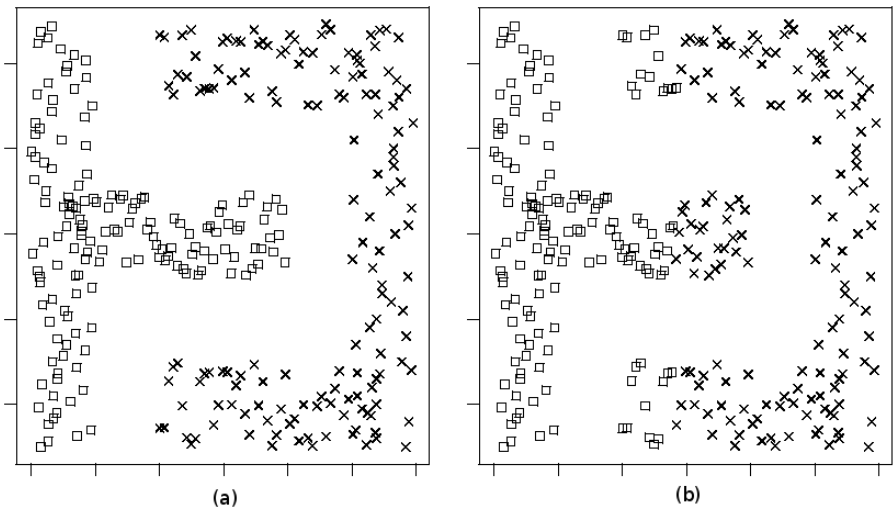

Figure 2: a) *T&U* dataset. A T-shaped cluster with the trunk in the concave part of a U-shaped cluster. b) The best partition that the k-means algorithm can find if k is set to 2.

In Figure 3a we present a 2-dimensional dataset with 4 spherical, compact and well separated clusters. No noise is present. In this kind of situation, the k-means algorithm can generally find the correct partition. We ran the k-means algorithm 10 times with $k = 4$ and in 8 cases the clusters were perfectly detected. Nevertheless in 2 cases the result was disappointing. This incorrect result can be seen in Figure 3b. Obviously, a 5 cluster partition, where the upper clusters are correctly split while the lower ones remain identical, would be a better partition.

With the previous examples we showed that well known and widely used algorithms can fail even when faced with spherical, compact and well separated clusters in a noiseless environment. Thus we consider that the problem is clearly illustrated.

4. Related work

Little theoretic work has been published about the evaluation of cluster validity indices (Bouguessa et al., 2006) and, as a consequence there is a lack of standard procedures to evaluate CVIs. These procedures are desirable for several reasons: comparisons of evaluations are currently unfeasible due to the heterogeneity of the published evaluations; each researcher must design its own procedures so the same work is repeated unnecessarily; incorrect procedures are designed based on intuitions and feelings...

The next example illustrates some of the effects of the lack of a theoretical framework in this area. Pal and Biswas (1997) compared 9 CVIs following the methodology explained in Section 2. They used 7 datasets in the experiment but

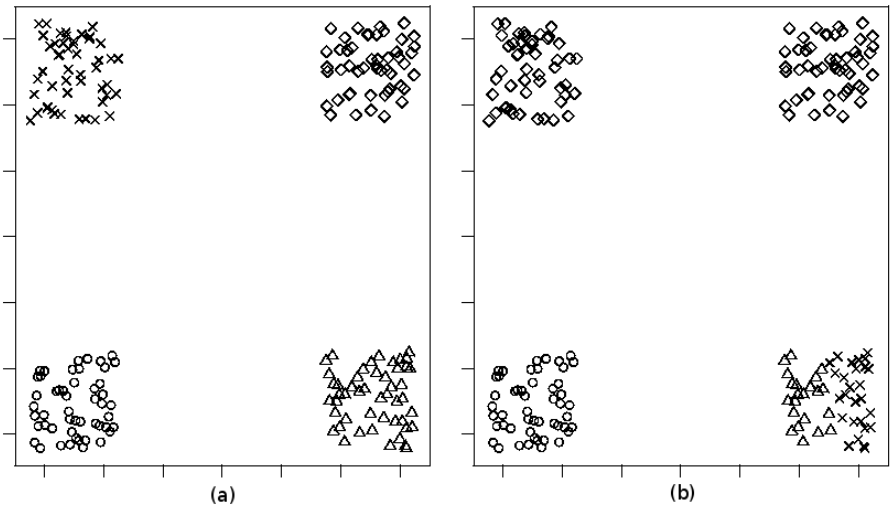

Figure 3: a) *Spheres* dataset. 4 compact and well separated clusters in a noiseless environment. b) An incorrect partition found by k-means due to an inappropriate centroid initialization.

they stated that in 2 of them "the cluster structure will be lost for all practical purposes". In any case, they treated these two datasets with no structure as the rest of the datasets. From our point of view, this is an incorrect procedure since we cannot expect a CVI to find a structure where no structure exists. We consider that if a CVI predicted a P^{nc} partition in this situation it was due to a lucky guess and it should not contribute to a better score for this CVI.

In spite of the need for standard procedures, as argued in previous paragraphs, most previous work has followed the classical methodology mentioned in Section 2. Some of them considered the problem arising from the algorithm correctness assumption, but others completely ignored it. Nevertheless, none of them proposed an objective and satisfactory solution to the problem. The following discussion focuses on how some previously published papers considered the algorithm correctness assumption.

Many authors did not even consider the algorithm correctness assumption and they carried out the experiment without checking if the P^{nc} partitions obtained were correct or were at least the best partitions the algorithms used could propose (Günter and Bunke, 2003; Kim and Ramakrishna, 2005; Maulik and Bandyopadhyay, 2002; Pal and Biswas, 1997). This type of work may be valid and the conclusions drawn from it may be correct, but we cannot be confident about the results obtained without a validation step.

There are some other studies where the algorithm correctness assumption was somehow considered. Hardy (1996) manually checked the partitions obtained and informed the reader about the correctness of these partitions. The author was aware of the fact that correct predictions were sometimes made based on incorrect partitions. This is one of the confusing effects of the algo-

rithm correctness assumption and, if ignored, can lead us to a misinterpretation of the results. Dimitriadou et al. (2002) were also conscious of the problem and they used multiple criteria to evaluate the indices. In any case, that work was restricted to binary datasets and the new proposed criteria seem to be useful just in this type of dataset. Chou et al. (2004) used algorithms sensitive to the initialization and repeated each execution several times. They computed each CVI for all the partitions obtained and considered in the results just the partition that was proposed most times. In this way they achieved a more robust experiment and some independence from the randomness of the clustering algorithm. In any case, this procedure cannot ensure that the algorithm correctness assumption holds. This is because there are some combinations of particular databases and algorithms where some of the non P^{nc} partitions, $\{P_i \in S | \, \mathrm{nc}(P^*) \neq \mathrm{nc}(P_i)\}$, generated by the algorithm fit better the data than most of the P^{nc} partitions. We have illustrated this fact in the example in Figure 2.

Milligan and Cooper (1985) carried out a wide CVI comparison; they compared 30 CVIs in a set of different environments. In spite of its age this work can be considered the most extensive and systematic CVI comparison. They implicitly considered the algorithm correctness assumption and performed a test to check if the assumption held. Milligan and Cooper combined many datasets and algorithms to build 432 different test solutions. They generated 25 partitions for every solution and computed each CVI for every partition. Before analysing the results, they performed a test step as explained below.

They compared all the partitions corresponding to a particular solution to the perfect partition of that solution using the Jaccard index (Jaccard, 1908) and the adjusted Rand index presented by Morey and Agresti (1984). Since they generated the databases synthetically and defined the clusters clearly, their perfect partition was known. The test step confirmed that for about 95% of the solutions the algorithm correctness assumption holds. Therefore, this value provides an upper limit to the CVIs evaluated. We consider that the remaining 5% should be ignored in the results analysis, since in those few cases the authors evaluated the CVIs under an incorrect assumption, but it is sound to think that the results would not vary qualitatively. We consider the high rate of "correct" clustering in the cited work is to be misleading since the datasets where composed of "internally cohesive and well separated" (Milligan and Cooper, 1985) clusters and the chosen algorithms fit the generated datasets very well. As argued in the introduction and shown in Section 7.2 we claim this rate would fall in real environments.

5. An alternative CVI evaluation methodology

Previous sections showed that the current CVI evaluation methodology will at best work in environments where the clustering algorithm is able to perfectly, or near perfectly, partition the data. In principle, this may not appear to be a problem since this behaviour can be induced and checked (Milligan and Cooper, 1985). However, we consider this procedure to be limited, since it only allows

evaluation of CVIs in situations where the clustering algorithm works under near optimal conditions: well structured data and some particular executions.

In real environments, it is usual to find datasets with overlapping and irregular clusters that make clustering algorithms work under far from optimal conditions. How well does a particular CVI behave in such a complex and irregular dataset? How does it behave when it is difficult for the clustering algorithm to find the structure underlying the data? These questions will remain unanswered with the classical methodology, but we consider them important questions. The evaluation of CVIs made under some particular conditions cannot be extrapolated to other conditions, so no CVI evaluation carried out previously can answer the questions posed above.

In this paper we propose an alternative CVI evaluation methodology, not based on the algorithm correctness assumption, that works correctly in every environment. Our proposal follows, to some extent, the idea found in Milligan and Cooper (1985): compare the generated partitions with the perfect partition using an external criterion. The underlying idea in the new methodology is to change the definition of a successful guess of a CVI. Instead of considering successful guesses to be the ones proposing the P^{nc} partition as best partition, we consider successful guesses to be the ones proposing the most similar partition to the perfect partition.

Both methodologies are very similar, but the new methodology needs one extra step. In this step the similarity between the m partitions in S and the perfect partition, P^*, must be computed. Let $\text{sim}(P_i, P_j)$ be a function that measures the similarity between partitions P_i and P_j. We call the partition obtaining the highest similarity value the most similar partition, $\hat{P}$.

$$\hat{P} = \arg\max_{P_i \in S}(\text{sim}(P^*, P_i))$$

Therefore, in the new methodology a successful guess will be the one that proposes the $\hat{P}$ partition as the best partition; that is, a successful guess exists if $\hat{P} = P^I$. The functions that measure the similarity between partitions are called in many different ways, such as external cluster validity indices or partition similarity measures. In the rest of the paper we will call them similarity measures as Pfitzner et al. (2009) do.

Notice that the only difference between the two methodologies is the definition of the target partition; that is, the partition the CVI must select as the best one. The target partition in the classical methodology is the one with the correct number of clusters, while the target partition in the new methodology is the one that most resembles the correct partition. Thus, if the most similar partition has the correct number of clusters, $\text{nc}(\hat{P}) = \text{nc}(P^{nc})$, then the target partition is the same for both methodologies, $\hat{P} \equiv P^{nc}$, and both methodologies are equivalent. Otherwise, the assumption does not hold and just the new methodology is able to capture the actual capabilities of the evaluated CVIs.

The similarity measure used in the experiment is a user-defined parameter of the methodology, so each experiment designer can choose the appropriate measure for his/her experiment. It is even possible to use several proximity

measures and combine their results by an averaging or voting process. In any case, the selection of a function to measure the similarity between two partitions is not a trivial task. Many have been proposed, but none of them is valid for all environments. (Meilă, 2005). More information about similarity measures can be found in (Albatineh et al., 2006; Batagelj and Bren, 1995; Baulieu, 1989; Pfitzner et al., 2009).

6. Experimental setup

In order to compare the two methodologies mentioned in this paper we performed an experiment where we ran the two CVI evaluation methodologies. In both methodologies the CVIs evaluated and the algorithm and datasets used were exactly the same. In this work we present results based on the VI similarity measure since it is theoretically well founded (Meilă, 2003, 2005). The VI similarity measure can be defined as:

$$\mathrm{VI}(P, P') = \mathrm{H}(P) + \mathrm{H}(P') - 2\mathrm{I}(P, P')$$

where H is the entropy of a partition, $\mathrm{H}(P) = -\sum_{C_i \in P} p(C_i) \log p(C_i)$ and I is the mutual information of two partitions, $\mathrm{I}(P, P') = \sum_{C_i \in P} \sum_{C'_i \in P'} p(C_i, C'_i) \frac{\log p(C_i, C'_i)}{p(C_i) p(C'_i)}$.

However, to assess to what extent the particular similarity measure used affects the results, we replicated the same experiment with 4 other partition similarity measures: Rand, Jaccard, Fowlkes-Mallows and a modified version of Rand (Jain and Dubes, 1988). These complementary similarity measures are described in Appendix A.

We evaluated 7 well known cluster validity indices on 7 synthetic 2-dimensional datasets and 3 datasets from real applications. We ran the k-means algorithm 10 times (with different random initializations) for each dataset and k parameter value. We set the k parameter to all values ranging from 2 to $\sqrt{n}$, where n represents the number of points in a particular dataset (Kim and Ramakrishna, 2005; Maulik and Bandyopadhyay, 2002).

6.1. Cluster validity indices

We mostly selected the CVIs evaluated from the indices examined in Milligan and Cooper (1985). We avoided the selection of indices that need a parameter or threshold value to be specified. We also discarded those indices specifically designed for hierarchical algorithms. Firstly, we selected all the indices that fulfilled these premises and were among the 5 indices considered best in Milligan and Cooper (1985): Calinski-Harabasz, C-Index and Gamma. We also analysed the G(+) and Davies-Bouldin indices, which are among the first 10 indices in Milligan and Cooper (1985). We also included in our work the McClain-Rao index which obtained one of the worst scores. Finally, we added to the group the widely used Dunn index (Chou et al., 2004; Maulik and Bandyopadhyay, 2002; Pal and Biswas, 1997). In the following paragraphs we give a brief description of the indices evaluated.

- Calinski-Harabasz (Calinski and Harabasz, 1974): This index is computed as

$$\frac{\text{trace}B/(k-1)}{\text{trace}W/(n-k)}$$

where n is the number of points in the dataset, k the number of clusters and B and W are the between and within cluster scatter matrices. The maximum value of the index is used to select the best partition.

- C-Index (Hubert and Levin, 1976): For this index, S, the sum of distances over all pairs of points from the same cluster, must be computed. Let n_w be the number of those pairs. S_{min} is the sum of the n_w smallest distances over all points in the dataset. Similarly S_{max} is the sum of the n_w largest distances. The C-Index is then defined as follows:

$$\frac{S - S_{min}}{S_{max} - S_{min}}$$

In this case the minimum value will denote the best partition.

- Gamma (Baker and Hubert, 1975): This index is an adaptation of Goodman and Kruskal's Gamma index. We define $\mathrm{d}_l(O_i, O_j)$ as the number of pairs of points in the dataset that are in different clusters and are more similar than O_i and O_j. The sum of all d_l values of all pairs of points from the same cluster must be computed. The sum is then normalized by dividing it by the maximum achievable value. The minimum value of the index denotes the best partition.

- G(+) (Rohlf, 1974): For this index all possible data quadruples (a, b, c, d) must be examined. We say that a quadruple is inconsistent (or disconcordant, Günter and Bunke, 2003) if one of the two following conditions is true:
  - $\mathrm{d}(a, b) < \mathrm{d}(c, d)$, a and b are in different clusters, and c and d are in the same cluster.
  - $\mathrm{d}(a, b) > \mathrm{d}(c, d)$, a and b are in the same cluster, and c and d are in different clusters.

  Here, $\mathrm{d}(i, j)$ denotes the distance between the i and j points. If $S(-)$ is the number of inconsistent quadruples and n_w is the number of within-cluster distances, the index is defined as follows:

$$\frac{2S(-)}{n_w(n_w - 1)}$$

The minimum value is used to select the best partition.

- Davies-Bouldin (Davies and Bouldin, 1979): This index is defined as follows:

$$\frac{1}{k}\sum_{i=1}^{k} \max_{j=1,\ldots,k;j\neq i}(d_{ij})$$

where

$$d_{ij} = \frac{s_i + s_j}{\mathrm{d}(c_i, c_j)}$$

In the formulae, k is the number of clusters, s_i is the average distance of all patterns in cluster i to their cluster centroid and $\mathrm{d}(c_i, c_j)$ is the distance between the centroids of clusters i and j. In this case the minimum value will denote the best partition.

- McClain-Rao (McClain and Rao, 1975): For this index, the ratio between the average within-cluster distance and the average between-cluster distance is computed for each cluster. The index is defined as the average of the individual cluster ratios. The minimum value of the index denotes the best partition.

- Dunn (Dunn, 1973): The Dunn index is computed as d_{min}/d_{max}, where d_{min} denotes the smallest distance between two points from different clusters and d_{max} the largest distance between two points from the same cluster. The maximum value will denote the best partition.

6.2. Datasets

We evaluated these indices on 10 datasets. We created 7 of these synthetically, so that we could control the number, shape, size, overlap, compactness and separation of the clusters. We generated 2-dimensional datasets to allow a visual check of the datasets and the partitions computed. To avoid any possible debate on the definition of the correct partition all the clusters are well defined and separated. We used 3 more datasets, based on real data, from the UCI repository (Asuncion and Newman, 2007). The following paragraphs describe these datasets.

The *Spheres* dataset, shown in Figure 3a, is the simplest dataset we used. It is composed of 4 spherical, compact and well separated clusters of 50 points each. The *Noisy* dataset is quite similar to the *Spheres* dataset (see Figure 1a). The difference is that we added 25 noise points following a uniform random distribution. The *Density* and *Size* datasets are also based on the *Spheres* dataset. The former differs from the *Spheres* in that two of the clusters are more compact than the other two; that is, the points are distributed over a less extensive area of the input space (see Figure 4a). In the latter, the four clusters are distributed in the same area they were in the *Spheres* dataset, but two of them are composed of 100 points while the remaining two are composed of just 25 points (see Figure 4b). Therefore, in the last 2 datasets, the symmetry found in the original dataset disappeared.

In Figure 5a we show the dataset named *3x3*. It is also composed of spherical, compact and well separated clusters, but in this case we defined 9 clusters in a regular mesh of 3 rows and 3 columns. Each cluster has 25 points. Figure 5b shows the *Arcs* dataset which contains 4 arc-shaped clusters (26 points each). The convex part of each cluster is located in the concave part of the cluster next to it. Finally, we used the *T&U* dataset shown in Figure 2a. This dataset has

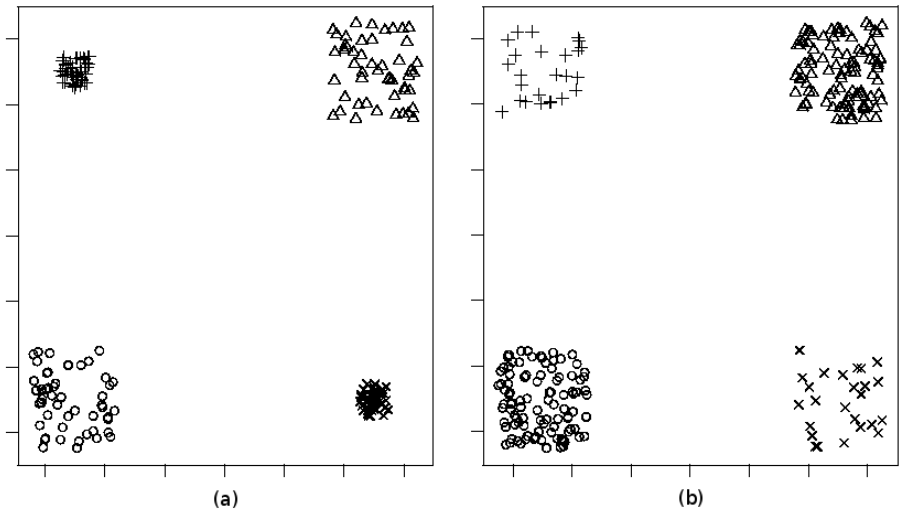

Figure 4: a) *Density* dataset. Four clusters with an identical number of points, but different densities. b) *Size* dataset. Four clusters distributed in the same area, but composed of a different number of points.

just two clusters of 150 points each. One of them is T-shaped while the other is U-shaped. The trunk of the T-shaped cluster is inside the concave part of the U-shaped one. Obviously, these two clusters are not linearly separable and the k-means algorithm will not be able to find the perfect partition (see Section 3).

In contrast, the real datasets have the following characteristics. The *Iris* dataset is widely used in clustering and validation problems. It has 4 attributes and 150 instances divided into 3 classes. The *Glass* dataset has 214 cases divided into 7 classes and 9 attributes describe each instance. Finally, the *Ecoli* dataset has 336 cases, 8 classes and 7 attributes.

As a summary, Table 1 shows the number of clusters and data points in each dataset.

7. Results

In this section we present the results obtained in both evaluations. We begin with the results we obtained with the classical methodology and continue with the ones obtained with the new methodology. Finally, we briefly analyse how the selected similarity measure affects the results for the alternative methodology.

It is not an easy task to select a simple and powerful representation of the results (Milligan and Cooper, 1985). In this paper we decided to use a simple tabular summary where we represent the number of successful guesses each CVI achieves for each of the datasets. This decision was motivated by the dependency of the CVIs on the datasets. This type of representation allows the examination

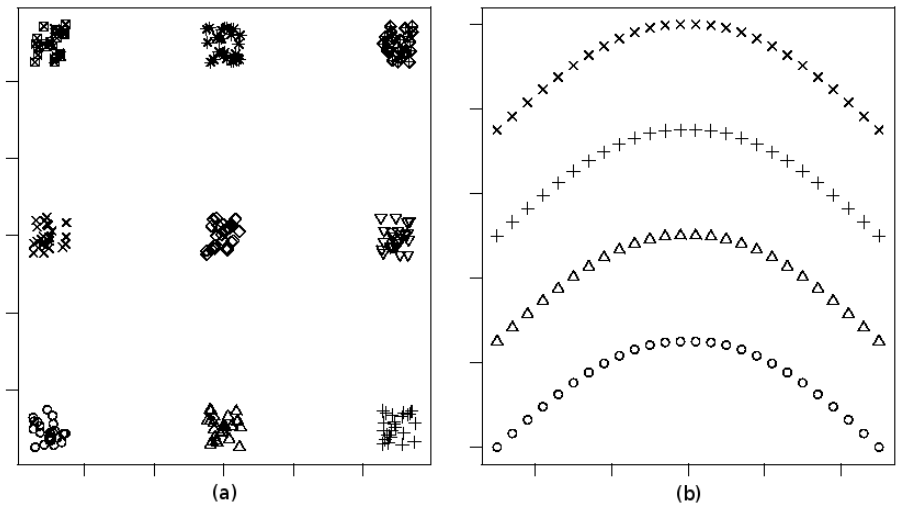

Figure 5: a) *3x3* dataset. 9 clusters distributed in a regular mesh of 3 rows and 3 columns. b) *Arcs* dataset. 4 arc shaped clusters.

of the overall behaviour of each CVI in different environments. Obviously, the meaning of "a successful guess" changes between the two methodologies.

7.1. Classical methodology

Table 2 summarizes the results for the classical methodology.

Based on these results, we could easily characterize the 7 cluster validity indices compared in this work. The Calinski-Harabasz index showed the best overall score. Surprisingly the C-Index, Gamma, and Dunn indices all showed the same results, not just for the overall score but also for each of the individual datasets. Moreover, a deeper analysis showed that the coincidence is complete, since all the successful guesses of these three indices occurred in exactly the same executions of the k-means algorithm. Thus, in this experiment the three indices mentioned were indistinguishable. The G(+) index would also be identical to those three indices if its 5 guesses in the *Ecoli* dataset were ignored.

The Calinski-Harabasz index also showed similar behaviour to the four grouped indices: we found differences in just three datasets. The most remarkable differences were in the *T&U* and *Iris* datasets. The Calinski-Harabasz index did a good job on these datasets while none of the other indices was able to make even one successful guess. On the other hand, its result on the *Density* dataset showed a lower score than the four grouped indices.

The remaining indices performed poorly. The Davies-Bouldin index obtained a score of 15, 14 of these in the *Spheres* and *Noisy* datasets. The McClain-Rao index obtained a significant 0 score.

On the other hand, the datasets also showed a particular characteristic. Most of the CVIs are able to do some satisfactory work with the 5 top datasets.

Dataset	Number of clusters	Number of data points
Spheres	4	200
Noisy	4	225
Density	4	200
Size	4	250
3x3	9	225
Arcs	4	104
T&U	2	300
Iris	3	150
Glass	7	214
Ecoli	8	336

Table 1: Characteristics of the datasets used in the study.

	Calinski-Harabasz	C-Index	Gamma	G(+)	Davies-Bouldin	McClain-Rao	Dunn
Spheres	8	8	8	8	8	0	8
Noisy	5	5	5	5	6	0	5
Density	2	5	5	5	1	0	5
Size	9	9	9	9	0	0	9
3x3	2	2	2	2	0	0	2
Arcs	0	0	0	0	0	0	0
T&U	7	0	0	0	0	0	0
Iris	6	0	0	0	0	0	0
Glass	0	0	0	0	0	0	0
Ecoli	0	0	0	5	0	0	0
Total	39	29	29	34	15	0	29

Table 2: Number of times each index proposes the P^{nc} partition as the best partition in each dataset.

Moreover, 90% of the successful guesses were found in those datasets. The remaining 5 datasets, which include all the real datasets, seemed to be quite difficult.

It is remarkable that the maximum achievable score was 100 (10 executions on 10 datasets) and even the best index obtained a score below 40%. We attributed these poor results to the fact that the k-means algorithm did not always define correct partitions and, therefore, the algorithm correctness assumption did not always hold. We check this hypothesis in the next section, where the same CVIs are evaluated using the proposed alternative methodology.

7.2. Alternative methodology

We must remember that both CVI evaluation methodologies are equivalent when the most similar partition to the correct partition is the one with the

correct number of clusters. Thus, before examining the results obtained with the alternative methodology, we checked how many times this occurred: for 32 of the 100 executions the $\hat{P}$ partition and the P^{nc} partition were the same. These results showed that the k-means algorithm was doing a far from perfect job and this also meant that, in this work, the scores for the new methodology could only reach an upper-bound of 68. Table 3 shows how these 68 differences are distributed between the datasets. The results for the *Spheres* and *Size* datasets must be similar, while the results for the *Arcs*, *T&U* and the 3 real datasets could show, although not necessarily, significant variations.

Spheres	*Noisy*	*Density*	*Size*	*3x3*	*Arcs*	*T&U*	*Iris*	*Glass*	*Ecoli*
2	5	5	1	7	9	9	10	10	10

Table 3: Number of executions of the k-means algorithm where the algorithm correctness assumptions does not hold.

Table 4 summarizes the results obtained in the evaluation performed with the alternative methodology. The scores are significantly higher. 5 of the indices

	Calinski-Harabasz	C-Index	Gamma	G(+)	Davies-Bouldin	McClain-Rao	Dunn
Spheres	10	10	10	9	8	0	8
Noisy	9	9	10	8	5	0	5
Density	4	8	10	8	3	0	5
Size	10	10	10	10	0	0	10
3x3	7	8	7	7	4	0	4
Arcs	8	0	0	0	0	0	1
T&U	1	0	0	0	0	0	0
Iris	4	9	10	0	10	0	10
Glass	5	7	7	0	5	0	7
Ecoli	0	5	1	0	0	0	8
Total	58	66	65	42	35	0	58

Table 4: Number of times each index proposes the $\hat{P}$ partition as the best partition in each dataset.

obtained a higher score than the highest score shown in Table 2 and the average score increased from 25.0% to 46.3%. This fact clearly shows that the cluster validity indices do a much better job of finding good partitions than finding the correct number of clusters. We consider the former must be the actual goal of a CVI, so we consider the results in Table 4 more significant than the results in Table 2.

In the following we analyse the results shown in Table 4 in a similar way as we did in the previous section and emphasize the differences in both sets of results. With respect to the ranking of indices the most remarkable facts were

three. First, the tie between Gamma, C-Index and Dunn was broken. The first two indices remained similar and obtained the top positions in the ranking. On the other hand, the improvements for the Dunn index were not enough to keep it together with Gamma and C-Index. Second, Calinski-Harabasz lost the first position and obtained the same score as the Dunn index. Nevertheless, the tie was arbitrary, since the distribution of guesses between the datasets was significantly different. Finally, the G(+) index, whose behaviour was similar to, and slightly better than, the behaviour of the 3 indices mentioned above, performed poorly with the new methodology. In particular, its 0 score on the real datasets put its overall score close to that of Davies-Bouldin.

Let us now focus on a more detailed analysis of the behaviour the indices showed. The C-Index showed a very regular improvement pattern. It obtained a better score in the alternative methodology in all but two datasets: *Arcs* and *T&U*. These two datasets, composed of non-convex clusters, proved to be extremely difficult for most of the indices. The results for Gamma were very similar to those obtained by C-Index. The main difference was found in the *Ecoli* dataset, where it obtained just a single correct guess. The Dunn index still showed a similar pattern to the two indices previously mentioned, but obtained lower scores. The main exception was the particularly good performance level it obtained for the *Ecoli* dataset. These three indices showed an impressive improvement for the three real datasets: from 0% to 71%.

We follow the analysis with the Calinski-Harabasz index, which obtained the same overall result as the Dunn index. However as stated previously, their behaviour differs significantly for each dataset. In general terms we can say that Calinski-Harabasz behaves better than the Dunn index for synthetic datasets, while the opposite occurs for real data. The Calinski-Harabasz index also showed us an interesting phenomenon that we call a "lucky guess". Notice that its score for the *T&U* and *Iris* datasets was lower in the alternative methodology than in the classical one. This is not an unusual event and, as noted by Hardy (1996), it means that the index did guess the correct number of clusters based on an incorrect partition.

The next index in the ranking was G(+). This index proved to be very similar to C-Index, Gamma and Dunn in the results for the classical methodology. The results for the alternative methodology showed that its performance on synthetic datasets remained at a good level, so its low position in the ranking was due to its 0 score for real datasets. Notice that its score for the *Ecoli* dataset was 5 for the classical methodology, which means these were "lucky guesses".

Finally, the worst two indices remained the same. Davies-Bouldin improved its results from 15 to 35, but this was not enough to move it higher in the ranking. McClain-Rao, once again, obtained a 0 score.

7.3. The effect of the similarity measure in the alternative methodology

In the follwing paragraphs we briefly discuss how the results obtained with the alternative methodology vary depending on the similarity measure used. As mentioned in the previous section, we replicated the experiment with Rand,

Jaccard, Fowlkes-Mallows and a modified version of Rand. The reader can find the detailed results in Appendix B.

The results showed that, regardless of the selected similarity measure, the CVI ranking obtained with the alternative methodology was different to the ranking obtained with the classical methodology. Furthermore, the score of the CVIs is significantly higher when the alternative methodology is used instead of the classical one. While the average success rate for the CVIs was 25% for the classical methodology, it ranged from 41% to 46% for the alternative methodology.

Next, we analysed how the CVIs were ranked by the alternative methodology depending on the used similarity measure. The results showed that Davies-Bouldin and McClain-Rao were the worst CVIs in every case. If we focus on the remaining five CVIs the results for Jaccard, Fowlkes-Mallows and modified Rand were very similar. All of them considered C-Index, Gamma and Calinski-Harabasz as the best CVIs with a success rate of about 60%. The score for G(+) and Dunn was about 45%.

The results obtained with the VI similarity measure, showed in Table 4, are similar to those mentioned in the previous paragraph. The main difference is the higher score for Dunn (58%) which makes it comparable to the best ranked CVIs: C-Index, Gamma and Calinski-Harabasz. Something similar occurs with G(+) if the Rand similarity measure is used, since it achieves a 55% score. In addition, the score of Calinski-Harabasz decreases to 49%, so, in this case, it is not clear whether it should be included in the top scoring group.

8. Conclusions

In this work we argued that the methodology that has been widely used to evaluate cluster validity indices was based on an assumption we have called the algorithm correctness assumption. We have proved that this assumption does not hold in many different environments, including widely used algorithms and datasets with compact and well separated clusters.

We proposed an alternative to this methodology, which does not depend on the algorithm correctness assumption and, thus, it overcomes the problems caused by it. The new methodology arises from the acceptance of the fact that, due to their limitations, clustering algorithms sometimes need to over- or under-partition the data in order to find a partition as similar as possible to the correct partition. Thus, the cluster validity indices must be evaluated on their ability to find partitions that resemble the perfect partition.

The analysis of the results obtained by applying the two methodologies to the same data and indices allow two main conclusions to be drawn. First, cluster validity indices do a much better job of finding partitions similar to the correct partition than finding the correct number of clusters (the average success rate increases from 25.0% to 46.3%). But, did the results also show a different ranking of the indices? Did the indices show different levels of dependency on the datasets? If no differences were found it could be argued that the classical methodology could be used instead, since the overall behaviour would not

change. Certainly, the classical methodology does not need to compute any partition similarity measure, thus, its computational cost is obviously lower, and it is preferable from this point of view. Nevertheless, we claim that the methodology we propose is based on more robust premises, as it does not depend on algorithm correctness and it better measures the capabilities of a CVI.

The second main conclusion confirmed that the score increment was not the only difference between the two methodologies: the evaluation of the indices differed depending on the methodology applied. For instance, the index that obtained the second best score with the classical methodology was beaten by 4 indices when we used the new methodology. In addition, the triple tie between C-Index, Gamma and Dunn disappeared in the new methodology.

We are aware that the datasets used affected the results. The score increments were different depending on the dataset. For one of them the score increment was just 7.14%. In contrast, another dataset achieved a 58.8% increment. The set of real datasets showed the most impressive improvement, increasing their average score from 5.2% to 41.9%. These results confirmed the need for some standard procedures for selecting the datasets.

In the new methodology the experiment designer must choose the similarity measure used to compare partitions. In this work, apart from the variation of information index, we used 4 more indices to compare partitions. All of them confirmed the two previously mentioned conclusions: the cluster validity indices obtain higher scores with the new methodology than with the classical one and besides, the CVI ranking differs for both methodologies. Therefore, the main conclusions of this work are confirmed by five well-known similarity measures.

However, in a CVI evaluation the choice of the similarity measure must be done with caution since results depend on this parameter. Nevertheless, it seems that, at least for the five similarity measures used in this work, the variation is small. In fact, three of the similarity measures showed a very similar behaviour while the other two mainly disagreed about the performance of one of the CVIs. In any case, to obtain robust results, we suggest to combine the results obtained by several similarity measures.

We also reviewed a set of previous studies evaluating cluster validity indices. A few of them somehow considered the algorithm correctness assumption, while most of them did not. We do not claim their conclusions are wrong, but we consider their results should be checked to asses their validity. However, the analysis of the work related to cluster validation led us to conclude that the lack of a standard methodology is not the only problem in the area. We also found that there is no agreement on the datasets and algorithms chosen; not even on the general properties they should have. For the datasets, there are many properties that will influence the results obtained: the number of dimensions, number of clusters, number of instances, the size and shape of the clusters... For the algorithms also, many different options can be chosen: partitional vs. hierarchical, hard vs. fuzzy, and so on. The way the parameters of parameterized CVIs must be defined is also an untreated issue (Chou et al., 2004; Dimitriadou et al., 2002; Hardy, 1996; Kim and Ramakrishna, 2005; Kothari and Pitts, 1999; Maulik and Bandyopadhyay, 2002; Milligan and Cooper, 1985; Pal and Biswas, 1997).

In this context we argue that a theoretical framework is needed in the area of cluster validation, and that standard procedures should be specified in order to achieve important improvements in this area. We claim that the methodology we have presented in this work should be used wherever a cluster validity index must be evaluated and that it should be part of these standard procedures.

Appendix A Definition of the complementary similarity measures

In this section we describe the four similarity measures used to replicate the experiment performed with the Variation of Information similarity measure.

Assuming that we must compare partitions P and P' we define a as the number of object pairs that belong to the same clusters in both partitions; b as the number of object pairs that belong to the same cluster in P, but not in P'; c as the number of object pairs that belong to the same cluster in P', but not in P; and d as the number of object pairs that belong to different clusters in both partitions. Based in Jain and Dubes (1988) we define the mentioned similarity measures as:

$$\text{Rand} = \frac{a+d}{a+b+c+d}$$

$$\text{Jaccard} = \frac{a}{a+b+c}$$

$$\text{Fowlkes} - \text{Mallows} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

$$\text{Modified Rand} = \frac{\left[(a+d)/\binom{n}{2}\right] - E\left[(a+d)/\binom{n}{2}\right]}{1 - E\left[(a+d)/\binom{n}{2}\right]}$$

Appendix B Results corresponding to the complementary similarity measures

The results obtained with the new methodology and the four complementary similarity measures are shown in Table B.1 (Rand), Table B.2 (Jaccard), Table B.3 (Fowlkes-Mallows) and Table B.4 (Modified Rand).

References

Albatineh, A. N., Niewiadomska-Bugaj, M., Mihalko, D., 2006. On similarity indices and correction for chance agreement. Journal of Classification 23, 301–313.

Asuncion, A., Newman, D., 2007. UCI machine learning repository. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`

Baker, F. B., Hubert, L. J., 1975. Measuring the power of hierarchical cluster analysis. Journal of the American Statistical Association 70, 31–38.

	Calinski-Harabasz	C-Index	Gamma	G(+)	Davies-Bouldin	McClain-Rao	Dunn
Spheres	10	10	10	9	8	0	8
Noisy	9	9	10	8	5	0	5
Density	4	8	10	8	3	0	5
Size	10	10	10	10	0	0	10
3x3	10	9	10	10	4	0	2
Arcs	0	6	7	5	6	4	3
T&U	0	0	0	0	0	0	0
Iris	6	0	0	0	0	0	0
Glass	0	0	0	5	0	4	0
Ecoli	0	4	4	0	0	0	4
Total	49	56	61	55	26	8	37

Table B.1: Number of times each index proposes the $\hat{P}$ partition (according to the Rand similarity measure) as the best partition in each dataset.

	Calinski-Harabasz	C-Index	Gamma	G(+)	Davies-Bouldin	McClain-Rao	Dunn
Spheres	10	10	10	9	8	0	8
Noisy	9	9	10	8	5	0	5
Density	4	8	10	8	3	0	5
Size	10	10	10	10	0	0	10
3x3	9	10	9	9	4	0	3
Arcs	9	0	0	0	0	0	1
T&U	0	0	0	0	0	0	0
Iris	6	0	0	0	0	0	0
Glass	2	10	10	0	2	0	8
Ecoli	0	6	3	0	0	0	6
Total	59	63	62	44	22	0	46

Table B.2: Number of times each index proposes the $\hat{P}$ partition (according to the Jaccard similarity measure) as the best partition in each dataset.

	Calinski-Harabasz	C-Index	Gamma	G(+)	Davies-Bouldin	McClain-Rao	Dunn
Spheres	10	10	10	9	8	0	8
Noisy	9	9	10	8	5	0	5
Density	4	8	10	8	3	0	5
Size	10	10	10	10	0	0	10
3x3	9	10	9	9	4	0	3
Arcs	9	0	0	0	0	0	1
T&U	0	0	0	0	0	0	0
Iris	6	0	0	0	0	0	0
Glass	3	9	9	0	3	0	7
Ecoli	0	5	1	0	0	0	8
Total	60	61	59	44	23	0	47

Table B.3: Number of times each index proposes the $\hat{P}$ partition (according to the Fowlkes-Mallows similarity measure) as the best partition in each dataset.

	Calinski-Harabasz	C-Index	Gamma	G(+)	Davies-Bouldin	McClain-Rao	Dunn
Spheres	10	10	10	9	8	0	8
Noisy	9	9	10	8	5	0	5
Density	4	8	10	8	3	0	5
Size	10	10	10	10	0	0	10
3x3	9	10	9	9	4	0	3
Arcs	6	1	1	1	1	1	1
T&U	0	0	0	0	0	0	0
Iris	6	0	0	0	0	0	0
Glass	2	6	6	0	2	1	4
Ecoli	0	5	4	0	0	0	5
Total	56	59	60	45	23	2	41

Table B.4: Number of times each index proposes the $\hat{P}$ partition (according to the Modified Rand similarity measure) as the best partition in each dataset.

Barbará, D., Jajodia, S. (Eds.), 2002. Applications of Data Mining in Computer Security. Kluwer Academic Publishers.

Batagelj, V., Bren, M., March 1995. Comparing resemblance measures. Journal of Classification 12 (1), 73–90.

Baulieu, F., 1989. A classification of presence/absence based dissimilarity coefficients. Journal of Classification 6 (1), 233–246.

Bouguessa, M., Wang, S., Sun, H., 2006. An objective approach to cluster validation. Pattern Recognition Letters 27 (13), 1419–1430.

Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics 3, 1–27.

Chou, C.-H., Su, M.-C., Lai, E., July 2004. A new cluster validity measure and its application to image compression. Pattern Analysis and Applications 7 (2), 205–220.

Davies, D. L., Bouldin, D. W., 1979. A clustering separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1, 224–227.

Devillez, A., Billaudelb, P., Lecolier, G. V., 2002. A fuzzy hybrid hierarchical clustering method with a new criterion able to find the optimal partition. Fuzzy Sets and Systems 128, 323–338.

Dimitriadou, E., Dolňicar, S., Weingessel, A., March 2002. An examination of indexes for determining the number of clusters in binary data sets. Psychometrika 67 (1), 137–159.

Dunn, J. C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3, 32–57.

Günter, S., Bunke, H., May 2003. Validation indices for graph clustering. Pattern Recognition Letters 24 (8), 1107–1113.

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2001. On clustering validation techniques. Journal of Intelligent Information Systems 17, 107–145.

Hardy, A., November 1996. On the number of clusters. Computational Statistics & Data Analysis 23 (1), 83–96.

Holzinger, K. J., Harman, H. H., 1941. Factor Analysis. Chicago: University of Chicago Press.

Hubert, L. J., Levin, J. R., 1976. A general statistical framework for assessing categorical clustering in free recall. Psychological Bulletin 83, 1072–1080.

Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. Bulletin de la Societé Vaudoise de Sciences Naturelles 44, 223–370.

Jain, A. K., Dubes, R. C., 1988. Algorithms for Clustering Data. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Kim, M., Ramakrishna, R. S., November 2005. New indices for cluster validity assessment. Pattern Recognition Letters 26 (15), 2353–2363.

Kothari, R., Pitts, D., April 1999. On finding the number of clusters. Pattern Recognition Letters 20 (4), 405–416.

Maulik, U., Bandyopadhyay, S., December 2002. Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (12), 1650–1654.

McClain, J. O., Rao, V. R., 1975. CLUSTISZ: A program to test for the quality of clustering of a set of objects. Journal of marketing research 12, 456–460.

Meilă, M., 2003. Comparing clusterings by the variation of information. In: Proceedings of the Sixteenth Annual Conference ofn Computational Learning Theory (COLT).

Meilă, M., 2005. Comparing clusterings – an axiomatic view. In: Proceedings of the Twenty-Second International Conference on Machine Learning (ICML 2005). pp. 577–584.

Milligan, G. W., Cooper, M. C., June 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50 (2), 159–179.

Mirkin, B., 2005. Clustering for Data Mining: A Data Recovery Approach. Chapman & Hall/CRC.

Morey, L. C., Agresti, A., 1984. The measurement of classification agreement: An adjustment to the rand statistic for chance agreement. Educational and Psychological Measurement 44, 33–37.

Pal, N. R., Biswas, J., June 1997. Cluster validation using graph theoretic concepts. Pattern Recognition 30 (6), 847–857.

Pfitzner, D., Leibbrandt, R., Powers, D., June 2009. Characterization and evaluation of similarity measures for pairs of clusterings. Knowledge and Information Systems 19 (3), 361–394.

Rohlf, F. J., 1974. Methods of comparing classifications. Annual Review of Ecology and Systematics 5, 101–113.

Sneath, P. H. A., Sokal, R. R., 1973. Numerical Taxonomy. Books in biology. W. H. Freeman and Company.