Pattern Recognition 46 (2013) 243–256

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

An extensive comparative study of cluster validity indices

Olatz Arbelaitz, Ibai Gurrutxaga*, Javier Muguerza, Jesús M. Pérez, Iñigo Perona

Department of Computer Architecture and Technology, University of the Basque Country UPV/EHU, Manuel Lardizabal 1, 20018 Donostia, Spain

ARTICLE INFO

Article history:
Received 3 February 2012
Received in revised form
26 July 2012
Accepted 31 July 2012
Available online 9 August 2012

Keywords:
Crisp clustering
Cluster validity index
Comparative analysis

ABSTRACT

The validation of the results obtained by clustering algorithms is a fundamental part of the clustering process. The most used approaches for cluster validation are based on internal cluster validity indices. Although many indices have been proposed, there is no recent extensive comparative study of their performance. In this paper we show the results of an experimental work that compares 30 cluster validity indices in many different environments with different characteristics. These results can serve as a guideline for selecting the most suitable index for each possible application and provide a deep insight into the performance differences between the currently available indices.

1. Introduction

Clustering is an unsupervised pattern classification method that partitions the input space into clusters. The goal of a clustering algorithm is to perform a partition where objects within a cluster are similar and objects in different clusters are dissimilar. Therefore, the purpose of clustering is to identify natural structures in a dataset [1–4] and it is widely used in many fields such as psychology [5], biology [4], pattern recognition [3], image processing [6] and computer security [7].

Once a clustering algorithm has processed a dataset and obtained a partition of the input data, a relevant question arises: How well does the proposed partition fit the input data? This question is relevant for two main reasons. First, an optimal clustering algorithm does not exist. In other words, different algorithms — or even different configurations of the same algorithm — produce different partitions and none of them have proved to be the best in all situations [8]. 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—frequently known as the k parameter. Since this information is rarely previously known, the usual approach is to run the algorithm several times with a different k value for each run. Then, all the partitions are evaluated and the partition that best fits the data is selected. The process of estimating how well a partition fits the structure underlying the data is known as cluster validation [1].

Cluster validation is a difficult task and lacks the theoretical background other areas, such as supervised learning, have. Moreover, a recent work argues the suitability of context-dependent evaluation methods [9]. Nevertheless, the authors also state that the analysis of cluster validation techniques is a valid research question in some contexts, such as clustering algorithms' optimization. Moreover, in our opinion, cluster validation tools analyzed in context-independent evaluations will greatly contribute to context-dependent evaluation strategies. Therefore, our work is based on a general, context-independent cluster evaluation process.

In this context, it is usual to classify the cluster validation techniques into three groups — internal, external and relative validation — but the classification criteria are not always clear [10,1,2,11]. In any case, there is a clear distinction between validation techniques if we focus on the information available in the validation process. Some techniques — related to external validation — validate a partition by comparing it with the correct partition. Other techniques — related to internal validation — validate a partition by examining just the partitioned data. Obviously, the former can only make sense in a controlled test environment, since in a real application the underlying structure of the data is unknown and, therefore, the correct partition is not available.

When the correct partition is available the usual approach is to compare it with the partition proposed by the clustering algorithm based on one of the many indices that compare data partitions; e.g. Rand, Adjusted Rand, Jaccard, Fowlkes–Mallows, Variation of Information [12].

On the other hand, when the correct partition is not available there are several approaches to validating a partition. One of them is to focus on the partitioned data and to measure the

* Corresponding author. Tel.: +34 943015166; fax: +34 943015590.
E-mail addresses: olatz.arbelaitz@ehu.es (O. Arbelaitz), i.gurrutxaga@ehu.es (I. Gurrutxaga), j.muguerza@ehu.es (J. Muguerza), txus.perez@ehu.es (J.M. Pérez), inigo.perona@ehu.es (I. Perona).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.07.021

compactness and separation of the clusters. In this case another type of index is used; e.g. Dunn [13], Davies–Bouldin [14], Calinski–Harabasz [15]. Another more recent approach is the stability based validation [16,17] which is not model dependant and does not require any assumption of compactness. This approach does not directly validate a partition, but it relies on the stability of the clustering algorithm over different samples of the input dataset.

The differences of the mentioned validation approaches make it difficult to compare all of them in the same framework. This work focuses on the first approach mentioned, which directly estimates the quality of a partition by measuring the compactness and separation of the clusters. Although there is no standard terminology, in the remainder of this paper we will call *Cluster Validity Index* (CVI) to these kind of indices. For the indices that compare two partitions we will use the term *Partition Similarity Measure*.

Previous works have shown that there is no single CVI that outperforms the rest [18–20]. This is not surprising since the same occurs in many other areas and this is why we usually deal with multiple clustering algorithms, partition similarity measures, classification algorithms, validation techniques, etc. This makes it obvious that researchers and practitioners need some guidelines on which particular tool they should use in each environment.

Focusing on internal cluster validation, we can find some works that compare a set of CVIs and, therefore, these could be used as guidelines for selecting the most suitable CVI in each environment. However, most of these comparisons are related to the proposal of a new CVI [6,21–24] or variants of known CVIs [25,8,26] and, unfortunately, the experiments are usually performed in restricted environments—few CVIs compared on few datasets, just one clustering algorithm implied. There are few works that do not propose a new CVI but compare some of them in order to draw some general conclusions [10,18,27,20]. Surprisingly, the 25 year-old paper of Milligan and Cooper [20] is the work most cited as a CVI comparison reference. Certainly, to the best of our knowledge, nobody has since published such an extensive and systematic comparative study.

In this paper we present the results of an extensive CVI comparison along the same lines as Milligan and Cooper [20], which is the last work that compares a set of 30 CVIs based on the results obtained in hundreds of environments. We claim that we have improved the referenced work in three main areas. First, we can compare many new indices that did not exist in 1985 and discard those that have rarely been used since. Second, we can take advantage of the increases in computational power achieved in recent decades to carry out a wider experiment. Finally, thanks to the advances in communication technologies we can easily store all the detailed results available in electronic format, so that every reader can access them and focus on the results that are relevant in his/her particular environment.

Moreover, our work is based on a corrected methodology that avoids an incorrect assumption made by the usual CVI comparison methodology [28]. Therefore, we present two main contributions in this paper. First, we present the main results of the most extensive CVI comparison ever carried out. Second, this comparison is the first extensive CVI comparison carried out with the methodological correction proposed by Gurrutxaga et al. [28]. Moreover, although the experiment's size prevents us from publishing all the results in this paper, they are all available in electronic format in the web.[1]

The next section discusses other works related to CVI comparison. Section 3 describes all the cluster validity indices compared in this work and Section 4 describes the particular details of the experimental design. In Section 5 we show the main results of the work and, finally, we draw some conclusions and suggest some possible extensions in Section 6

2. Related work

Most of the works that compare CVIs use the same approach: A set of CVIs is used to estimate the number of clusters in a set of datasets partitioned by several algorithms. The number of successes of each CVI in the experiment can be called its score and is considered an estimator of its "quality". For a more formal description of this methodology and a possible alternative to it see [28].

Despite this widely used approach, most of the works are not comparable since they differ in the CVIs compared, datasets used, results analysis. In this section we overview some of the works that compare a set of CVIs, focusing on the experiment characteristics.

The paper published by Milligan and Cooper [20] in 1985 is still the work of reference on internal cluster validation. That work compared 30 CVIs. The authors called them "Stopping criteria" because they were used to stop the agglomerative process of a hierarchical clustering algorithm [2,4] and this is why the experiments were done with hierarchical clustering algorithms (single-linkage, complete-linkage, average-linkage and Ward). They used 108 synthetic datasets with a varying number of non-overlapped clusters (2, 3, 4 or 5), dimensionality (4, 6 or 8) and cluster sizes. They presented the results in a tabular format, showing the number of times that each CVI predicted the correct number of clusters. Moreover, the tables also included the number of times that the prediction of each CVI overestimated or underestimated the real number of clusters by 1 or 2.

The same tabular format was used by Dubes [27] two years later. The novelty of this work is that the author used some tables where the score of each CVI was shown according to the values of each experimental factor—clustering algorithm, dataset dimensionality, number of clusters. Moreover, he used the χ^2 statistic to test the effect of each factor on the behaviour of the compared CVIs. Certainly, the use of statistical tests to validate the experimental results is not common practice in clustering, as opposed to other areas such as supervised learning. The main drawback of this work is that it compares just 2 CVIs (Davies–Bouldin and the modified Hubert statistic). The experiment is performed in 2 parallel works of 32 and 64 synthetic datasets, 3 clustering algorithms (single-linkage, complete-linkage and CLUSTER) and 100 runs. The datasets' characteristics were controlled in the generation process and they used different sizes (50 or 100 objects), dimensionality (2, 3, 4 or 5), number of clusters (2, 4, 6 or 8), sampling window (cubic or spherical) and cluster overlap.

In 1997, Bezdek et al. [29] published a paper comparing 23 CVIs based on 3 runs of the EM algorithm and 12 synthetic datasets. The datasets were formed by 3 or 6 Gaussian clusters and the results were presented in tables that showed the successes of every CVI on each dataset. Another work that compared 15 CVIs was performed by Dimitriadou et al. [18] based on 100 runs of k-means and hard competitive learning algorithms. The 162 datasets used in this work were composed of binary attributes which made the experiment and the results presentation somewhat different to the previously mentioned ones.

More recently, Brun et al. [10] compared 8 CVIs using several clustering algorithms: k-means, fuzzy c-means, SOM, single-linkage, complete-linkage and EM. They used 600 synthetic datasets based on 6 models with varying dimensionality (2 or 10), cluster shape (spherical or Gaussian) and number of clusters (2 or 4). The novelty in this work can be found in the

[1] http://www.sc.ehu.es/aldapa/cvi.

comparison methodology. The authors compared the partitions obtained by the clustering algorithms with the correct partitions and computed an error value for each partition. Then, the “quality” of the CVI is measured as its correlation with the measured error values. In this work, not just internal but also external and relative indices are examined. The results show that the Rand index is highly correlated with the error measure.

The mentioned correlation between the error measure and the Rand index makes one think about the adequacy of the error as a definitive measure. In the recent work of Gurrutxaga et al. [28] the authors accepted that there is no single way of establishing the quality of a partition and they proposed using one of the external indices available—or even better, several of them. This is the first work that clearly confronted a methodological drawback ignored by many authors, but noticed by others [10,22,23,20]. Since the main goal of this work was to present a modification of the traditional methodology, they compared just 7 CVIs based on 7 synthetic and 3 real datasets and 10 runs of the k-means algorithm.

Other CVI comparisons can be found where new CVIs are proposed, but in this case the experiment is usually limited. It is common to find works comparing 5 or 10 CVIs on a similar number of datasets [6,21,22,25,8,26,24].

3. Cluster validity indices

In this section we describe the 30 CVIs compared in this work. First, to simplify and reduce the CVI description section we define the general notation used in this paper and particular notations used to describe several indices.

3.1. Notation

Let us define a dataset X as a set of N objects represented as vectors in an F-dimensional space: $X=\{x_1,x_2,\ldots,x_N\}\subseteq\Re^F$. A partition or clustering in X is a set of disjoint clusters that partitions X into K groups: $C=\{c_1,c_2,\ldots,c_K\}$ where $\bigcup_{c_k\in C}c_k=X, c_k\cap c_l=\emptyset\ \forall k\neq l$. The centroid of a cluster c_k is its mean vector, $\overline{c_k}=1/|c_k|\sum_{x_i\in c_k}x_i$ and, similarly, the centroid of the dataset is the mean vector of the whole dataset, $\overline{X}=1/N\sum_{x_i\in X}x_i$.

We will denote the Euclidean distance between objects x_i and x_j as $\mathrm{d_e}(x_i,x_j)$. We define the Point Symmetry-Distance [30] between the object x_i and the cluster c_k as

$$\mathrm{d}^*_{\mathrm{ps}}(x_i,c_k)=1/2\sum\min(2)_{x_j\in c_k}\{\mathrm{d_e}(2\overline{c_k}-x_i,x_j)\}.$$

The point $2\overline{c_k}-x_i$ is called the symmetric point of x_i with respect to the centroid of c_k. The function $\sum\min$ can be seen as a variation of the min function where $\sum\min(n)$ computes the sum of the n lowest values of its argument. Similarly, we can define the $\sum\max$ function as an analogue variation of the max function.

Finally, let us define n_w since it is used by several indices. n_w is the number of object pairs in a partition that are in the same cluster, $n_w=\sum_{c_k\in C}\binom{|c_k|}{2}$.

3.2. Index definitions

Next, we describe the 30 CVIs compared in this work. We focused on CVIs that can be easily evaluated by the usual methodologies and avoided those that could lead to confusion due to the need for a subjective decision by the experimenter. Therefore, we have discarded some indices that needed to determine a “knee” in a plot — such as the Modified Hubert index [31] — need to tune a parameter or need some kind of normalization — such as the υ_{SV} index [32] or the Jump index [33]. We have also avoided fuzzy indices, since our goal was to focus on crisp clustering. In brief, we focused on crisp CVIs that allow selection of the best partition based on its lowest or highest value.

Most of the indices estimate the cluster cohesion (within or intra-variance) and the cluster separation (between or inter-variance) and combine them to compute a quality measure. The combination is performed by a division (ratio-type indices) or a sum (summation-type indices) [25].

For each index we define an abbreviation that will be helpful in the results section. Moreover, we accompanied each abbreviation with an up or down arrow. The down arrow denotes that a lower value of that index means a “better” partition. The up arrow means exactly the opposite.

- Dunn index (D↑) [13]: This index has many variants and some of them will be described next. It is a ratio-type index where the cohesion is estimated by the nearest neighbour distance and the separation by the maximum cluster diameter. The original index is defined as

$$\mathrm{D}(C)=\frac{\min_{c_k\in C}\{\min_{c_l\in C\setminus c_k}\{\delta(c_k,c_l)\}\}}{\max_{c_k\in C}\{\Delta(c_k)\}},$$

where

$$\delta(c_k,c_l)=\min_{x_i\in c_k}\min_{x_j\in c_l}\{\mathrm{d_e}(x_i,x_j)\},$$

$$\Delta(c_k)=\max_{x_i,x_j\in c_k}\{\mathrm{d_e}(x_i,x_j)\}.$$

- Calinski–Harabasz (CH ↑) [15]: This index obtained the best results in the work of Milligan and Cooper [20]. It is a ratio-type index where the cohesion is estimated based on the distances from the points in a cluster to its centroid. The separation is based on the distance from the centroids to the global centroid, as defined in Section 3.1. It can be defined as

$$\mathrm{CH}(C)=\frac{N-K}{K-1}\frac{\sum_{c_k\in C}|c_k|\mathrm{d_e}(\overline{c_k},\overline{X})}{\sum_{c_k\in C}\sum_{x_i\in c_k}\mathrm{d_e}(x_i,\overline{c_k})}.$$

- Gamma index (G ↓) [34]: The Gamma index is an adaptation of Goodman and Kruskal’s Gamma index and can be described as

$$\mathrm{G}(C)=\frac{\sum_{c_k\in C}\sum_{x_i,x_j\in c_k}\mathrm{dl}(x_i,x_j)}{n_w\left(\binom{N}{2}-n_w\right)},$$

where $\mathrm{dl}(x_i,x_j)$ denotes the number of all object pairs in X, namely x_k and x_l, that fulfil two conditions: (a) x_k and x_l are in different clusters, and (b) $\mathrm{d_e}(x_k,x_l)<\mathrm{d_e}(x_i,x_j)$. In this case the denominator is just a normalization factor.

- C-Index (CI↓) [35]: This index is a type of normalized cohesion estimator and is defined as

$$\mathrm{CI}(C)=\frac{\mathrm{S}(C)-\mathrm{S}_{min}(C)}{\mathrm{S}_{max}(C)-\mathrm{S}_{min}(C)},$$

where

$$\mathrm{S}(C)=\sum_{c_k\in C}\sum_{x_i,x_j\in c_k}\mathrm{d_e}(x_i,x_j),$$

$$\mathrm{S}_{min}(C)=\sum\min(n_w)_{x_i,x_j\in X}\{\mathrm{d_e}(x_i,x_j)\},$$

$$\mathrm{S}_{max}(C)=\sum\max(n_w)_{x_i,x_j\in X}\{\mathrm{d_e}(x_i,x_j)\}.$$

- Davies–Bouldin index (DB↓) [14]: This is probably one of the most used indices in CVI comparison studies. It estimates the cohesion based on the distance from the points in a cluster to its centroid and the separation based on the distance between centroids. It is defined as

$$\mathrm{DB}(C)=\frac{1}{K}\sum_{c_k\in C}\max_{c_l\in C\backslash c_k}\left\{\frac{\mathrm{S}(c_k)+\mathrm{S}(c_l)}{\mathrm{d_e}(\overline{c_k},\overline{c_l})}\right\},$$

where

$$\mathrm{S}(c_k)=1/|c_k|\sum_{x_i\in c_k}\mathrm{d_e}(x_i,\overline{c_k}).$$

- Silhouette index (Sil↑) [36]: This index is a normalized summation-type index. The cohesion is measured based on the distance between all the points in the same cluster and the separation is based on the nearest neighbour distance. It is defined as

$$\mathrm{Sil}(C)=1/N\sum_{c_k\in C}\sum_{x_i\in c_k}\frac{\mathrm{b}(x_i,c_k)-\mathrm{a}(x_i,c_k)}{\max\{\mathrm{a}(x_i,c_k),\mathrm{b}(x_i,c_k)\}},$$

where

$$\mathrm{a}(x_i,c_k)=1/|c_k|\sum_{x_j\in c_k}\mathrm{d_e}(x_i,x_j),$$

$$\mathrm{b}(x_i,c_k)=\min_{c_l\in C\backslash c_k}\left\{1/|c_l|\sum_{x_j\in c_l}\mathrm{d_e}(x_i,x_j)\right\}.$$

- Graph theory based Dunn and Davies–Bouldin variations (D^{MST}↑, D^{RNG}↑, D^{GG}↑, DB^{MST}↓, DB^{RNG}↓, DB^{GG}↓) [8]: These indices are variations of Dunn and Davies–Bouldin. The variation affects how the cohesion estimators are computed—$\Delta(c_k)$ for the Dunn index and $\mathrm{S}(c_k)$ for the Davies–Bouldin index.
For each of the three versions — MST, RNG and GG — these two functions are computed in the same way. First, a particular type of graph is computed for c_k, taking the objects in the cluster as vertices and the distance between objects as the weight of each edge. Then the largest weight is taken as the value for $\Delta(c_k)$ and $\mathrm{S}(c_k)$. The difference between the three variants comes from the selected graph type. For MST a Minimum Spanning Tree is built, for RNG a Relative Neighbourhood Graph and for GG a Gabriel Graph.
- Generalized Dunn indices gD31↑, gD41↑, gD51↑, gD33↑, gD43↑, gD53↑) [37]: All the variations are a combination of three variants of δ — separation estimator — and two variations of Δ — cohesion estimator. Actually, Bezdek and Pal [37] proposed 6×3 variants — including the original index — but we selected those proposals that showed the best results. Therefore we analyzed the variants 3, 4 and 5 for δ and 1 and 3 for Δ.

$$\delta^3(c_k,c_l)=\frac{1}{|c_k||c_l|}\sum_{x_i\in c_k}\sum_{x_j\in c_l}\mathrm{d_e}x_i,x_j,$$

$$\delta^4(c_k,c_l)=\mathrm{d_e}(\overline{c_k},\overline{c_l}),$$

$$\delta^5(c_k,c_l)=\frac{1}{|c_k|+|c_l|}\left(\sum_{x_i\in c_k}\mathrm{d_e}(x_i,\overline{c_k})+\sum_{x_j\in c_l}\mathrm{d_e}(x_j,\overline{c_l})\right)$$

and

$$\Delta^1(c_k)=\Delta(c_k),$$

$$\Delta^3(c_k)=2/|c_k|\sum_{x_i\in c_k}\mathrm{d_e}(x_i,\overline{c_k}).$$

- S_Dbw index (SDbw↓) [38]: This is a ratio-type index that has a more complex formulation based on the Euclidean norm $\|x\|=(x^Tx)^{1/2}$, the standard deviation of a set of objects, $\sigma(X)=1/|X|\sum_{x_i\in X}(x_i-\overline{x})^2$ and the standard deviation of a partition, $\mathrm{stdev}(C)=1/K\sqrt{\sum_{c_k\in C}\|\sigma(c_k)\|}$. Its definition is

$$\mathrm{SDbw}(C)=1/K\sum_{c_k\in C}\frac{\|\sigma(c_k)\|}{\|\sigma(X)\|}+\frac{1}{K(K-1)}\sum_{c_k\in C}\sum_{c_l\in C\backslash c_k}\frac{\mathrm{den}(c_k,c_l)}{\max\{\mathrm{den}(c_k),\mathrm{den}(c_l)\}},$$

where

$$\mathrm{den}(c_k)=\sum_{x_i\in c_k}f(x_i,\overline{c_k}),$$

$$\mathrm{den}(c_k,c_l)=\sum_{x_i\in c_k\cup c_l}f\left(x_i,\frac{\overline{c_k}+\overline{c_l}}{2}\right),$$

and

$$\mathrm{f}(x_i,c_k)=\begin{cases}0 & \text{if } \mathrm{d_e}(x_i,\overline{c_k})>\mathrm{stdev}(C),\\ 1 & \text{otherwise.}\end{cases}$$

- CS index (CS↓) [6]: This index was proposed in the image compression environment, but can be extended to any other environment. It is a ratio-type index that estimates the cohesion by the cluster diameters and the separation by the nearest neighbour distance. Its definition is

$$\mathrm{CS}(C)=\frac{\sum_{c_k\in C}\{1/|c_k|\sum_{x_i\in c_k}\max_{x_j\in c_k}\{\mathrm{d_e}(x_i,x_j)\}\}}{\sum_{c_k\in C}\min_{c_l\in C\backslash c_k}\{\mathrm{d_e}(\overline{c_k},\overline{c_l})\}}.$$

- Davies–Bouldin* (DB*↓) [25]: This variation of the Davies–Bouldin index was proposed together with an interesting discussion about different types of CVIs. Its definition is

$$\mathrm{DB}^*(C)=1/K\sum_{c_k\in C}\frac{\max_{c_l\in C\backslash c_k}\{\mathrm{S}(c_k)+\mathrm{S}(c_l)\}}{\min_{c_l\in C\backslash c_k}\{\mathrm{d_e}(\overline{c_k},\overline{c_l})\}}.$$

- Score function (SF↑) [39]: This is a summation-type index where the separation is measured based on the distance from the cluster centroids to the global centroid and the cohesion is based on the distance from the points in a cluster to its centroid. It is defined as

$$\mathrm{SF}(C)=1-\frac{1}{e^{e^{\mathrm{bcd}(C)+\mathrm{wcd}(C)}}},$$

where

$$\mathrm{bcd}(C)=\frac{\sum_{c_k\in C}|c_k|\mathrm{d_e}(\overline{c_k},\overline{X})}{N\times K},$$

$$\mathrm{wcd}(C)=\sum_{c_k\in C}1/|c_k|\sum_{x_i\in c_k}\mathrm{d_e}(x_i,\overline{c_k}).$$

- Sym-index (Sym↑) [30]: This index is an adaptation of the I index [19] based on the Point Symmetry-Distance. It is defined as

$$\mathrm{Sym}(C)=\frac{\max_{c_k,c_l\in C}\{\mathrm{d_e}(\overline{c_k},\overline{c_l})\}}{K\sum_{c_k\in C}\sum_{x_i\in c_k}\mathrm{d}^*_{\mathrm{ps}}(x_i,c_k)}.$$

- Point Symmetry-Distance based indices (SymDB↓, SymD↑, Sym33↑) [26]: These three indices are also based on the Point Symmetry-Distance and modify the cohesion estimator of the Davies–Bouldin, Dunn and generalized-Dunn (version 33) indices.

The SymDB index is computed as DB, but the computation of S is redefined as follows:

$$\mathrm{S}(c_k) = 1/|c_k| \sum_{x_i \in c_k} \mathrm{d}^*_{\mathrm{ps}}(x_i, c_k).$$

The symD index is like D, but the Δ function is defined as

$$\Delta(c_k) = \max_{x_i \in c_k}\{\mathrm{d}^*_{\mathrm{ps}}(x_i, c_k)\}.$$

And finally, the Sym33 index is a modification of gD33 where Δ is defined as

$$\Delta(c_k) = 2/|c_k| \sum_{x_i \in c_k} \mathrm{d}^*_{\mathrm{ps}}(x_i, c_k).$$

- COP index (COP↓) [40]: Although this index was first proposed to be used in conjunction with a cluster hierarchy postprocessing algorithm, it can also be used as an ordinary CVI. It is a ratio-type index where the cohesion is estimated by the distance from the points in a cluster to its centroid and the separation is based on the furthest neighbour distance. Its definition is

$$\mathrm{COP}(C) = \frac{1}{N}\sum_{c_k \in C} |c_k| \frac{1/|c_k| \sum_{x_i \in c_k} \mathrm{d_e}(x_i, \overline{c_k})}{\min_{x_i \notin c_k} \max_{x_j \in c_k} \mathrm{d_e}(x_i, x_j)}.$$

- Negentropy increment (NI↓) [23]: This is an index based on cluster normality estimation and, therefore, is not based on cohesion and separation estimations. It is defined as

$$\mathrm{NI}(C) = 1/2 \sum_{c_k \in C} p(c_k)\log|\Sigma_{c_k}| - 1/2\log|\Sigma_X| - \sum_{c_k \in C} p(c_k)\log p(c_k)$$

where $p(c_k) = |c_k|/N$, Σ_{c_k} denotes the covariance matrix of cluster c_k, Σ_X denotes the covariance matrix of the whole dataset and $|\Sigma|$ denotes the determinant of a covariance matrix. Although the authors proposed the index as defined above, they later proposed a correction due to the poor results obtained. Nevertheless, we will use the index in its original form since the correction does not meet the CVI selection criterion used for this work.

- SV-Index (SV↑) [24]: This ratio-type index is one of the most recent CVIs compared in this work. It estimates the separation by the nearest neighbour distance and the cohesion is based on the distance from the border points in a cluster to its centroid. It is defined as

$$\mathrm{SV}(C) = \frac{\sum_{c_k \in C} \min_{c_l \in C \setminus c_k}\{\mathrm{d_e}(\overline{c_k}, \overline{c_l})\}}{\sum_{c_k \in C} 10/|c_k| \sum \max_{x_i \in c_k}(0.1|c_k|)\{\mathrm{d_e}(x_i, \overline{c_k})\}}.$$

- OS-Index (OS↑) [24]: This is another recent ratio-type index proposed by Žalik and Žalik [24] where a more complex separation estimator is used. It is defined as

$$\mathrm{OS}(C) = \frac{\sum_{c_k \in C} \sum_{x_i \in c_k} \mathrm{ov}(x_i, c_k)}{\sum_{c_k \in C} 10/|c_k| \sum \max_{x_i \in c_k}(0.1|c_k|)\{\mathrm{d_e}(x_i, \overline{c_k})\}},$$

where

$$\mathrm{ov}(x_i, c_k) = \begin{cases} \dfrac{\mathrm{a}(x_i, c_k)}{\mathrm{b}(x_i, c_k)} & \text{if } \dfrac{\mathrm{b}(x_i, c_k) - \mathrm{a}(x_i, c_k)}{\mathrm{b}(x_i, c_k) + \mathrm{a}(x_i, c_k)} < 0.4, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\mathrm{a}(x_i, c_k) = 1/|c_k| \sum_{x_j \in c_k} \mathrm{d_e}(x_i, x_j),$$

$$\mathrm{b}(x_i, c_k) = 1/|c_k| \sum_{x_j \notin c_k} \min(|c_k|)\{\mathrm{d_e}(x_i, x_j)\}.$$

4. Experimental setup

In this section we describe the experiment performed to compare the CVIs listed in the previous section. As shown in Section 2, there are many possible experimental designs for such a comparison. Since we want to compare the CVIs in a wide variety of configurations we designed an experiment with several factors. Unfortunately, due to combinatorial explosion we had to limit each factor to just a few levels and this finally led us to an experiment with 6480 configurations.

The comparative methodology that we used is a variation of the traditional problem of estimating the number of clusters of a dataset. The usual approach is to run a clustering algorithm over a dataset with a set of different values for the k parameter — the number of clusters of the computed partition — obtaining a set of different partitions. Then, the evaluated CVI is computed for all the partitions. The number of clusters in the partition obtaining the best results is considered the prediction of the CVI for that particular dataset. If this prediction matches the true number of clusters, the prediction is considered successful.

The variation we used modifies the problem so that the CVIs are not used to estimate the correct number of clusters. They are used to predict which is the "best" partition in the mentioned set of partitions. The "best" partition is defined as the one that is the most similar to the correct one—measured by a partition similarity measure—which is not always the one with the correct number of clusters. For a formal and more detailed description see [28]. In order to avoid the possible bias introduced by the selection of a particular partition similarity measure, we replicated all the experiments using three partition similarity measures: Adjusted Rand [31], Jaccard [41] and Variation of Information [42].

We used three clustering algorithms to compute partitions from the datasets: k-means, Ward and average-linkage [2]. These are well known and it is easy to obtain different partitions by modifying the parameter that controls the number of clusters of the output partition. Each algorithm was used to compute a set of partitions with the number of clusters ranging from 2 to $\sqrt{N}$, where N is the number of objects in the dataset. In the case of the real datasets, the number of clusters in a partition was limited to 25 to avoid computational problems with large datasets.

As usual, we used several synthetically generated datasets for the CVI evaluation. Furthermore, we also compared them using 20 real datasets drawn from the UCI repository [43]. In any case, it is important to note that results based on real datasets should be analyzed with caution since these datasets are usually intended to be used with supervised learning and, therefore, they are not always well adapted to the clustering problem [9]. On the contrary, the synthetic datasets avoid many problems found with real datasets. For instance, in synthetic datasets categories exists independent of human experience and their characteristics can be easily controlled by the experiment designer.

The synthetic datasets were created to cover all the possible combinations of five factors: number of clusters (K), dimensionality (dim), cluster overlap (ov), cluster density (den) and noise level (nl). We defined two types of overlap: *strict*, meaning that the ov overlap level must be exactly satisfied, and *bounded*, meaning that ov is the maximum allowed overlap.

A fixed hypercubic sampling window is defined to create all the synthetic datasets. The window is defined by the $(0,0,\ldots,0)$ and $(50,50,\ldots,50)$ coordinates. In a similar way, a reduced sampling

Table 1
Values of the parameters used in the synthetic dataset generation step.

Param.	Value
n_{min}	100
K	2, 4, 8
dim	2, 4, 8
ov	1.5 (strict), 5 (bounded)
den	1, 4
nl	0, 0.1

window is defined by the (3,3,...,3) and (47,47,...,47) coordinates. Then, the centre for the first cluster, c_0, is randomly drawn in the reduced sampling window based on a uniform distribution. The first cluster is created by randomly drawing $n_{min} \times den$ points following a multivariate normal distribution of *dim* dimensions with mean c_0 and the identity as covariance matrix. All points located outside the sampling window are removed and new points are drawn to replace them.

The remaining clusters will have n_{min} points and this produces a density asymmetry when $den \neq 1$. This occurs because a different number of points will be located in the same approximate volume.

In particular, we build the remaining $K-1$ clusters as follows: if the overlap is *bounded*, the centre of the cluster, c_i, is drawn uniformly from the reduced sampling window. Otherwise, a previously created cluster centre, c_k, is randomly selected and the new cluster centre, c_i, is set to a random point located at a distance of $2 \times ov$ from c_k. In any case, if $d_e(c_i, c_l) < 2 \times ov \ \forall c_l \neq c_i$ the cluster centre is discarded and a new one is selected. Once the cluster centre has been defined the cluster is built by drawing n_{min} points in the same way as we did for the first cluster.

Finally, when all the clusters have been built, $nl \times N'$ points are randomly created following a uniform distribution in the sampling window, where N' is the number of non-noise points in the dataset, $N' = n_{min} \times (den + K - 1)$.

The values of the parameters used to create the synthetic datasets are shown in Table 1, making 72 different configurations. As we created 10 datasets from each configuration we used 720 synthetic datasets. Multiplying this value by three partition similarity measures and three clustering algorithms we obtain the 6480 configurations previously mentioned. Notice that the n_{min} parameter ensures that every cluster is composed of at least 100 objects.

Fig. 1 shows an example of 4 two-dimensional datasets we have used. In the figure we can see how the different values of the generation parameters affects the point distribution in the datasets. Fig. 1a shows a dataset with four clusters, with no cluster overlap, no noise and no density asymmetry. The other three plots show dataset with similar characteristics except for overlap, density and noise parameters.

The 20 real datasets and their main characteristics are shown in Table 2. In this case the experiment is based on 180 configurations—20 datasets, 3 algorithms and 3 partition similarity measures.

Including synthetic and real datasets, and taking into account the different number of partitions computed for each dataset, each of the 30 CVIs was computed for 156 069 partitions.

5. Results

One of the goals of this work is to present the results in such a way that readers can focus on the particular configurations they are interested in. However, the vast amount of results obtained prohibits all of them being shown in this paper. Therefore, we focus here on the overall results; drawing some important conclusions. However, all the detailed results are available in the web.

In this section we first describe the results obtained for the synthetic datasets and then the results for the real datasets are described. Finally, we present a brief discussion on the use of statistical analysis in clustering and we show the conclusions we drew by applying some statistical tests to the results.

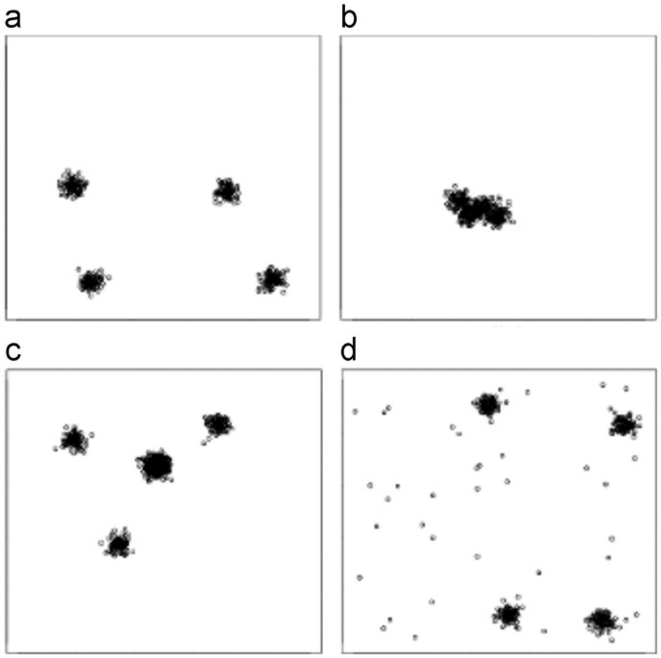

Fig. 1. Two-dimensional plots of four synthetic datasets used in the experiment. (a) Shows a "neutral" dataset with no cluster overlap, no density asymmetry and no noise. (b) Shows a similar dataset with high cluster overlap. (c) Shows a dataset with cluster density asymmetry. (d) Shows a dataset with noise.

Table 2
The characteristics of the real datasets drawn from the UCI repository.

Dataset	No. of objects	Features	Classes
Breast tissue	106	9	6
Breast Wisconsin	569	30	2
Ecoli	336	7	8
Glass	214	9	7
Haberman	306	3	2
Ionosphere	351	34	2
Iris	150	4	3
Movement libras	360	90	15
Musk	476	166	2
Parkinsons	195	22	2
Segmentation	2310	19	7
Sonar all	208	60	2
Spectf	267	44	2
Transfusion	748	4	2
Vehicle	846	18	4
Vertebral column	310	6	3
Vowel context	990	10	11
Wine	178	13	3
Winequality red	1599	11	6
Yeast	1484	8	10

5.1. Synthetic datasets

The overall results for the synthetic datasets are shown in Fig. 2. The figure shows the percentage of correct guesses

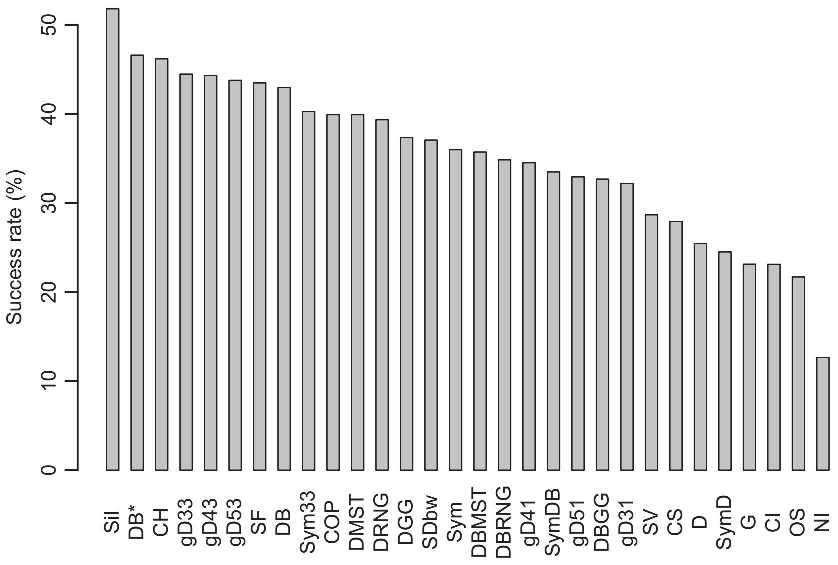

Fig. 2. Overall results for the experiment with synthetic datasets.

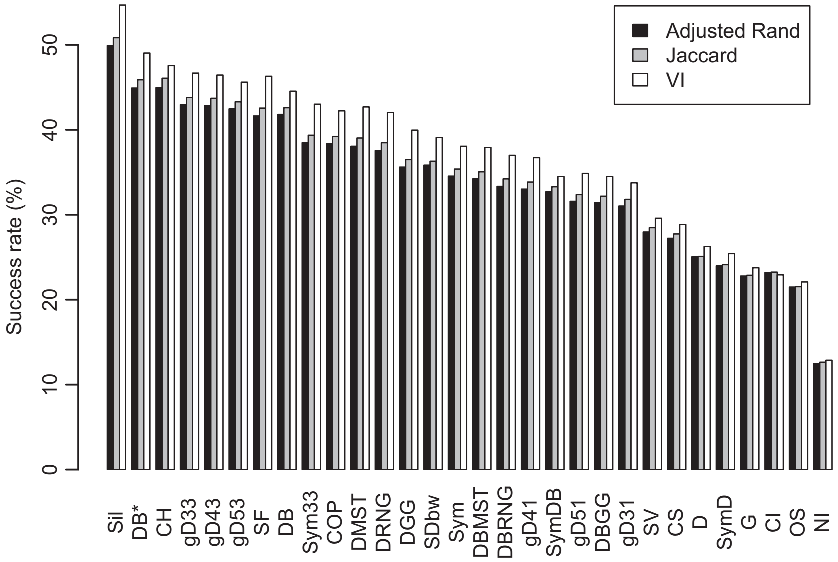

Fig. 3. Results for synthetic datasets broken down by partition similarity measure.

(successes) achieved by each CVI, which are sorted by the number of successes. Notice that this percentage refers to the 6480 configurations. The graph shows that Silhouette achieves the best overall results and is the only one that exceeds the 50% score. DB* and CH also show a good result, with a success rate beyond 45%.

It is also noticeable that in most cases variations of a CVI behave quite similarly; they appear in contiguous positions in the figure. The clearest cases are the generalized Dunn indices that use Δ^3 as cohesion estimator — gD33, gD43 and gD53 — and the graph theory based Dunn indices — D^{MST}, D^{RNG} and D^{GG}.

Next we will show a similar graph for each experimental factor. In this case the value of each CVI is shown for each value of the analyzed factor. We will keep the CVI order shown in Fig. 2, so a decreasing graph will denote that the analyzed factor does not change the overall ranking.

First of all, let us focus on the graph corresponding to the partition similarity measure. Remind that this is a parameter of the validation methodology we have used (see Section 4). In Fig. 3 we can see that the selected partition similarity measure does not affect the results. This result suggests that the CVI comparison is not affected by the particular selection of a parameter of the evaluation methodology and, therefore, we can be confident of the results. Also notice that although Adjusted Rand and Jaccard show very similar results the use of the VI partition similarity measure produces slightly higher success rates.

In the following figures a similar breakdown can be found with regard to the characteristics of the datasets. In Fig. 4 we can see how the number of clusters of the datasets affects the results. As expected, all the CVIs obtain better results with fewer clusters—average result for $k=2$ drops from 50.2% to 30.7% ($k=4$) and 24.8% ($k=8$). We can also see that for high values of this parameter the differences between the CVIs are reduced. Furthermore, some indices, such as COP, show little sensitivity to this parameter making it the best CVI for $k=8$.

With respect to dimensionality (see Fig. 5), the results show that the difficulty imposed by an increment in the number of dimensions does not severely affect the behaviour of the CVIs—except for NI. Moreover, some indices, such as Sym, show

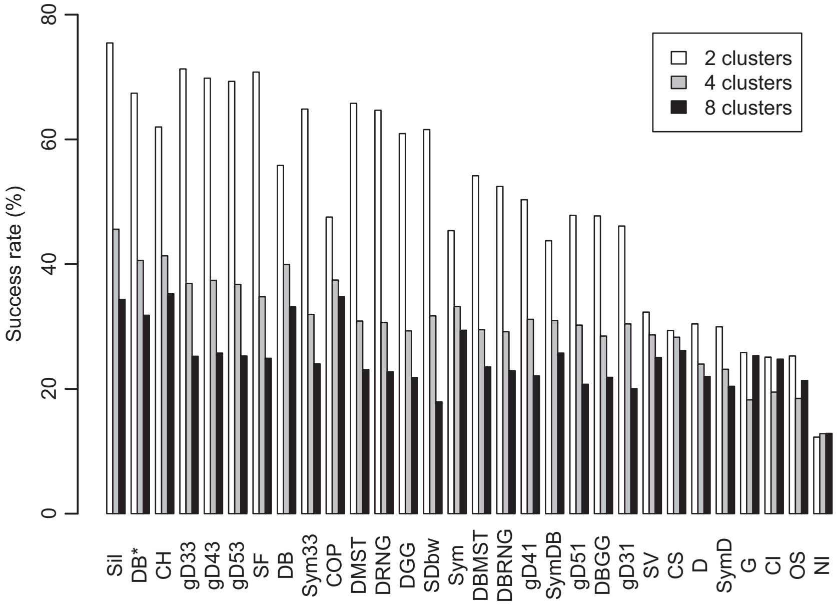

Fig. 4. Results for synthetic datasets broken down by number of clusters.

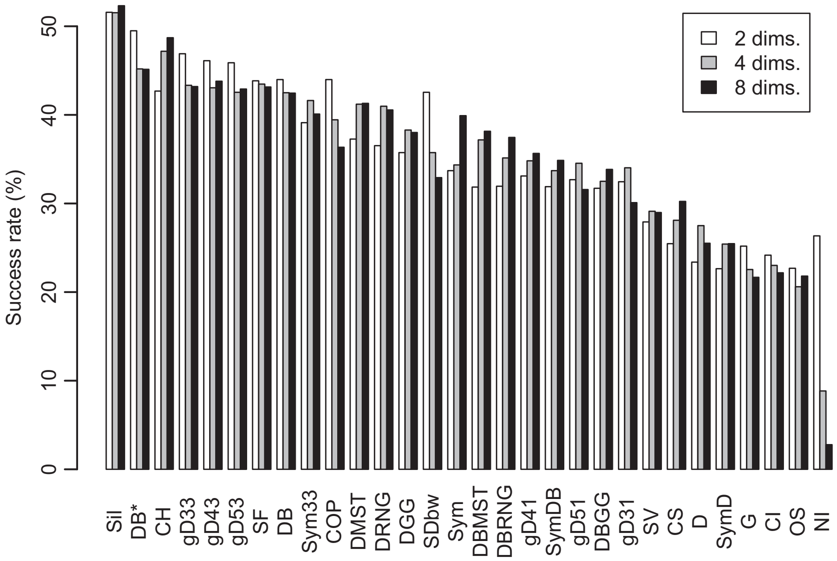

Fig. 5. Results for synthetic datasets broken down by dimensionality.

a better behaviour for datasets with higher dimensionality. Silhouette also achieves the best results for every value analyzed for this parameter.

Let us now focus on the results shown in Fig. 6. This graph shows that, as expected, datasets with no overlapping clusters lead to better CVI success rates. The average result decreases from 52.9% to 17.6% when well separated clusters are replaced by overlapped clusters. The graph also shows that although this parameter does not severely affect the overall trend, some CVIs are more hardly affected by cluster overlap, e.g. DB, COP and SymDB. Some others, such as G, CI and OS, seem not to work at all when clusters overlap.

With respect to the density of the clusters, Fig. 7 shows that having a cluster four times denser than the others, does not severely affect the CVIs. It seems that the best behaving indices are quite insensitive to this parameter while the rest show a better result when density heterogeneity is present. Silhouette is again, clearly, the CVI showing best results.

Noise level, the last dataset characteristic analyzed in this work, has a major impact on the scores of the CVIs (Fig. 8). In fact, the scores in noisy environments are on average three times lower than they are when no noise is present. Silhouette, and mostly SDbw, are the main exception to this rule since they show similar score values for noisy and noiseless environments. Besides, the overall trend is not always followed and CH is the CVI that achieves the best results when no noise is present.

Finally, Fig. 9 shows how the clustering algorithm used in the experiment affects the scores of the indices evaluated. Although we cannot find a clear pattern, it seems that the overall comparative results are not severely affected since the decreasing pattern of the graph is somehow maintained. Most of the CVIs obtain their worst results for the k-means algorithm, but there are some exceptions where the opposite holds—COP, G, CI and OS are the most remarkable examples. Silhouette is again the one achieving the best results for hierarchical algorithms, but CH is the best CVI when k-means is used as clustering algorithm.

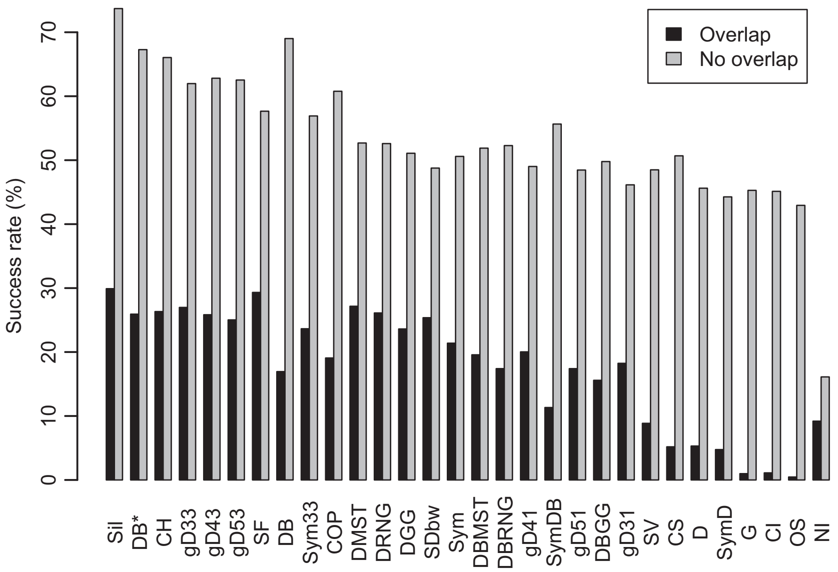

Fig. 6. Results for synthetic datasets broken down by cluster overlap.

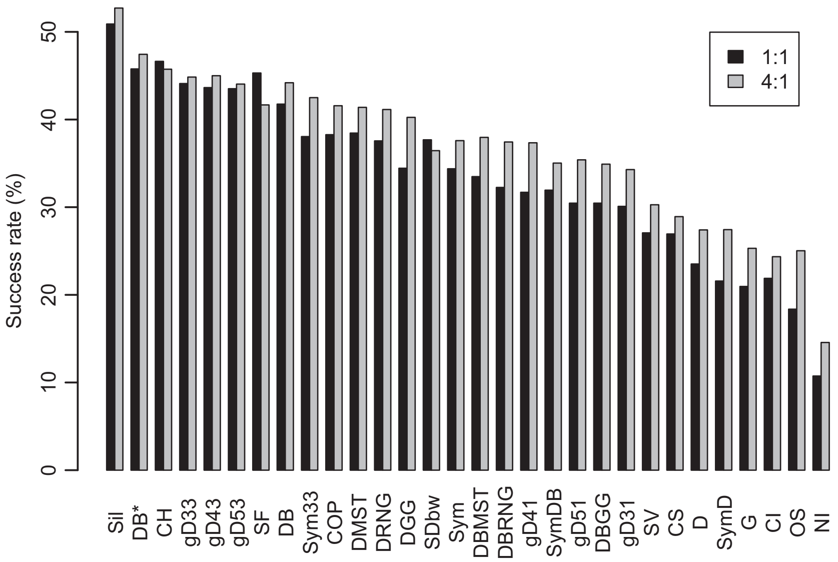

Fig. 7. Results for synthetic datasets broken down by density.

5.2. *Real datasets*

In this section we show the results obtained for 20 real datasets following a similar style to the one we used for synthetic datasets. Obviously, since we do not have control over the dataset design the number of experimental factors is reduced to 2: partition similarity measure and clustering algorithms.

First, in Fig. 10 we show the overall results for real datasets. A quick comparison to the overall results for the synthetic datasets (Fig. 2) shows that the results are qualitatively similar. Most of CVIs that obtained worst results with synthetic datasets are also in the tail of the ranking in the figure for real datasets. Focusing on the head of the ranking we can see that the generalized Dunn indices — gD33, gD43 and gD53 — remain in a similar position; SF, graph theory based Dunn and COP improve their position; and Silhouette, DB* and CH go down the ranking. Considering these results we can say that the mentioned generalizations of the Dunn index show the steadiest results.

Returning to the two experimental factors involved in the experiments with real datasets, in Fig. 11 we show the results broken down by partition similarity measure. We can see that in this case it seems that the partition similarity measure selected can affect the results. Although Jaccard and VI follow the overall pattern the Adjusted Rand index does not. Furthermore, it is clear that in every case the average scores are much lower when Adjusted Rand is used, dropping from 39.1% (VI) or 31.1% (Jaccard) to 10.0%.

With regard to the clustering algorithm used (see Fig. 12) the results are contradictory. On the one hand, if we focus on k-means and Ward, it seems that this factor does not severely affect the results. On the other hand, results for average-linkage reduce the differences between CVIs and do not follow the overall results. In this case, Sym shows the best results while SF achieves the highest success rates for k-means and Ward.

5.3. *Statistical tests*

Although the assessment of the experimental results using statistical tests is a widely studied technique in machine learning, it is rarely used in the clustering area. Among the works cited in

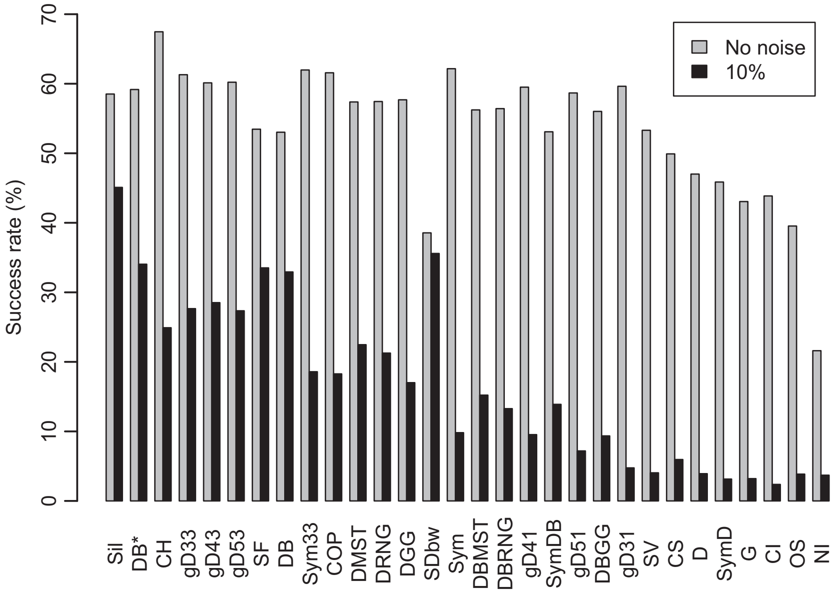

Fig. 8. Results for synthetic datasets broken down by noise.

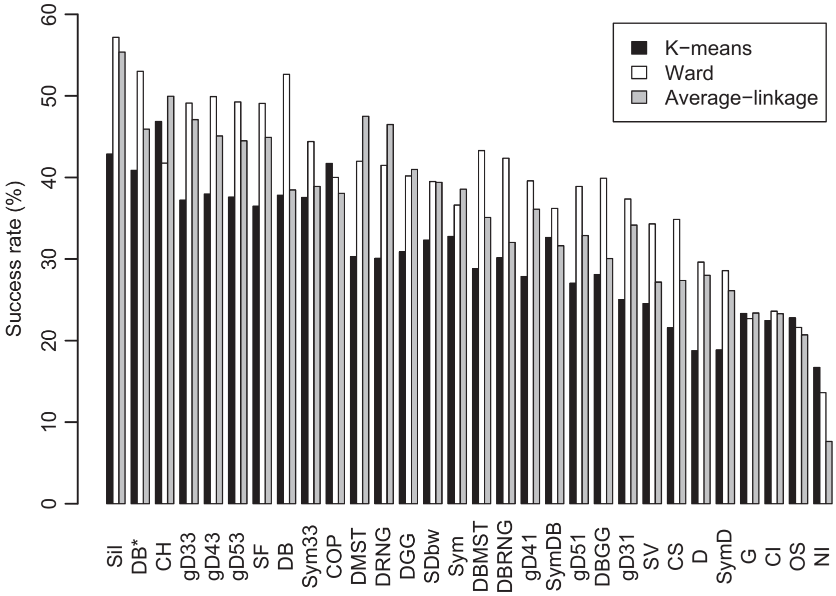

Fig. 9. Results for synthetic datasets broken down by clustering algorithm.

Section 2 just Dubes [27] used a statistical test to assess the influence of each experimental factor on the results obtained. However, in our case we focused on checking whether the observed differences between CVIs were statistically significant or not.

We argue that an effort should be made by the clustering and statistics communities to adapt these tools to clustering and effectively introduce them in the area. These types of tests would be even more important in extensive comparative works such as the one described in this paper. Therefore, although it is not the goal of this work, we propose a possible direct adaptation of a comparison method used in supervised learning. This method has been chosen due to the proximity of the supervised learning area to clustering and because the use of statistical tests in this area has been widely studied [44–46].

We next describe the method and the proposed adaptation. Then, we conclude this section by discussing the results obtained when we applied the proposed tests to the results obtained in the experiment carried out in our work to compare the performance of CVIs.

We based our statistical method on a common scenario in supervised learning where classification algorithms are compared. In this case it is usual to run the algorithms on several datasets and to compute a ‘‘quality’’ estimate, such as the accuracy or the AUC value, for each algorithm and database pair. A usual approach is to test the quality values achieved by all the algorithms for each dataset independently [45]. However, Demšar [44] recently argued that a single test based on all the algorithms and all the datasets is a better choice. One of the advantages of this method is that the different values compared in the statistical test are independent, since they come from different datasets.

We have adapted the method proposed by Demšar [44] and subsequently extended by García and Herrera [46] to CVI comparisons. In brief, we simply replaced the classification algorithms

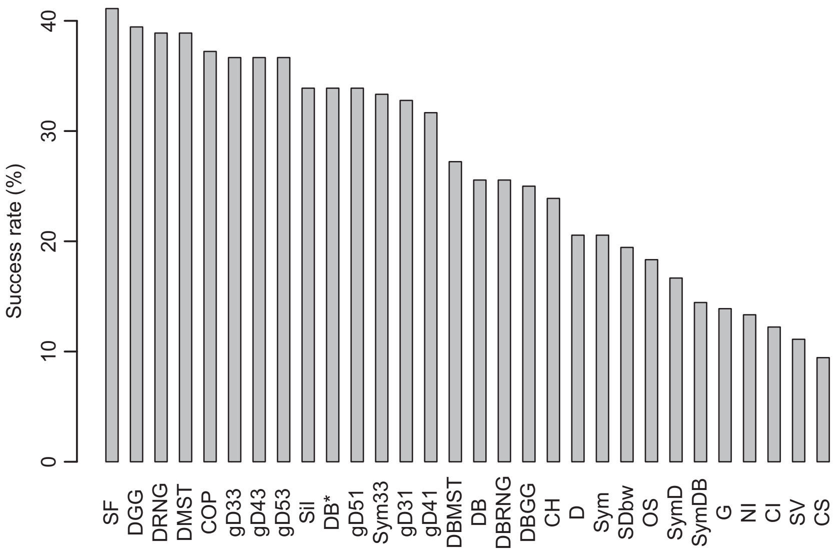

Fig. 10. Overall results for real datasets.

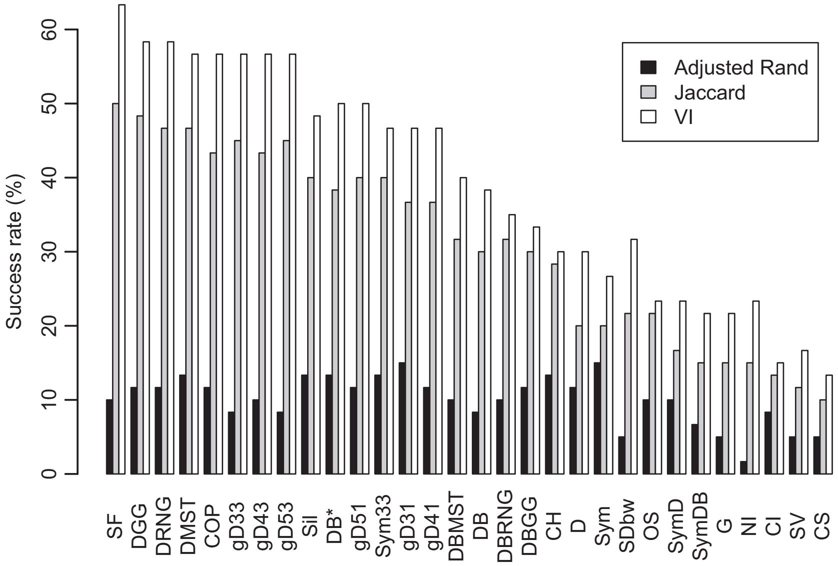

Fig. 11. Results for real datasets broken down by partition similarity measure.

by CVIs. However, this is not enough, since in our experiments we obtained a Boolean value for each CVI-configuration pair instead of a "quality" estimate. Moreover, the configurations we obtained by varying the clustering algorithm and partition similarity measure are based on the same dataset, so it can be argued that they are not sufficiently independent.

Our solution was to add for each dataset the number of successes each CVI obtained for each clustering algorithm–partition similarity measure pair. Moreover, in order to obtain a more precise estimate, we also added the number of successes obtained in every run—remember that we created 10 datasets for each combination of dataset characteristics. We thus obtained 72 values ranging from 0 to 90 for each CVI, that gave us a "quality" estimate for independent datasets. Finally, we applied the statistical tests with no further modifications.

The tests we used were designed for comparisons of multiple classifiers (CVIs) in an all-to-all way. We used the Friedman test to check if any statistical difference existed and the Nemenyi test for pairwise CVI comparison [44]. Furthermore, we performed additional pairwise CVI comparisons with the Shaffer test as suggested by García and Herrera [46]. In both cases we performed the tests with 5% and 10% confidence level.

The main conclusion obtained by applying the above tests is that there are undoubtedly statistically significant differences between the 30 CVIs, as the Friedman test categorically shows with a p-value on the order of 10^{-80}. All the performed pair-wise comparisons show a very similar result, so in Fig. 13 we only show the results for the most powerful test that we performed—Shaffer with a confidence level of 10%.

Since the used statistical tests are based on average rank values, the figure shows all the CVIs sorted by average rank. The results are very similar to those based on average scores (Fig. 2), but there are a couple of differences that should be underlined. First of all, the CVI order slightly changed, but most of the movements occurred in the central part of the ranking. Secondly, the CVIs formed quite well separated groups. In the first group there are 10 indices with an average rank between 9 and 13. Taking into account variations of a CVI as a single one, the group contains six indices: Silhouette,

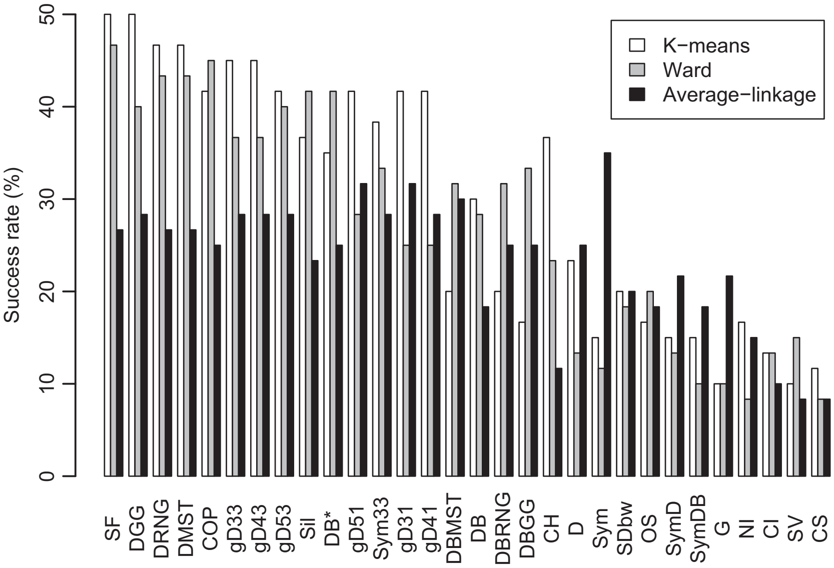

Fig. 12. Results for real datasets broken down by clustering algorithm.

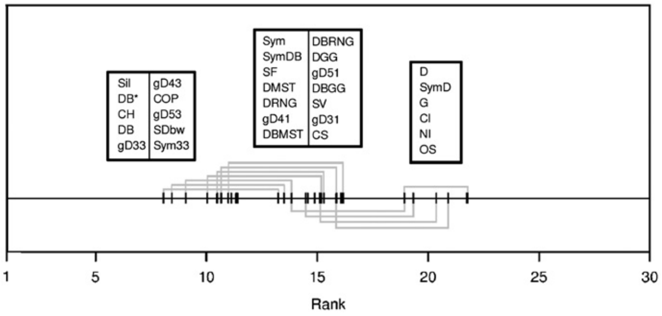

Fig. 13. Results for Shaffer test with a significance level of 10%.

Davies–Bouldin, Calinski–Harabasz, generalized Dunn, COP and SDbw. There is also a crowded central group with 14 CVIs and average rank between 14 and 17; and finally, a group of six indices with average rank between 19 and 23.

The bars in the figure group the indices that do not show statistically significant differences. The highly overlapped bars difficult the task of drawing categorical conclusions, but on the following we resume the information in the graph and remark the most interesting points:

- No significant difference exists between CVIs in the same group.
- All the CVIs in the first group perform significantly better than the CVIs in the third group.
- The best behaving CVI, Sil, obtains significantly better results than all the CVIs in the second group, except Sym.
- All the CVIs in the second group, except Sym and SymDB, have no statistically significant differences with at least one CVI in the third group.

In conclusion, the data does not show sufficiently strong evidence to distinguish a small set of CVIs as being significantly better than the rest. Nevertheless, there is a group of about 10 indices that seems to be recommendable and Silhouette, Davies–Bouldin* and Calinski–Harabasz are in the top of this group. We have also performed statistical test to the experiment subsets shown in the results section, but no CVI can be considered significantly better than the others in any case.

6. Conclusions and further work

In this paper we presented a comparison of 30 cluster validity indices on an extensive set of configurations. It is, to the best of our knowledge, the most extensive CVI comparison ever published. Moreover, it is the first non-trivial CVI comparison that uses the methodological correction recently proposed by Gurrutxaga et al. [28].

Due to the huge size of the experiment we have not been able to show all the results obtained. However, the interested reader can access them in electronic format in the web. The great advantage of this is that readers can focus on the results for the configurations they are interested in and we therefore provide a tool to enable them to select the most suitable CVIs for their particular application. This procedure is very recommendable since there is not a single CVI that showed clear advantage over the rest in every context, although Silhouette index obtained the best results in many of them.

We next summarize the main conclusions we drew from the CVI comparison. First, we observed that some CVIs appear to be more suitable for certain configurations, although the results were not conclusive. Furthermore, the overall trend never changed dramatically when we focused on a particular factor. Another fact worth noting is that the results for real and synthetic datasets are qualitatively similar, although they show disagreements for some particular indices.

With regard to the experimental factors, noise and cluster overlap had the greatest impact on CVI performance. The number of successes is dramatically reduced when noise is present or clusters overlap. In particular, the inclusion of 10% random noise reduces the average score to a third part. A very similar score reduction was found when the clusters were moved closer so they highly overlapped. Another remarkable and surprising fact is that some indices showed better results in (a priori) more complex configurations. For example, some indices improved their results when the dimensionality of the datasets increased or the homogeneity of the cluster densities disappeared.

Finally, we confirmed that the selection of a partition similarity measure that enables correction of the experimental methodology is not a critical factor. Nevertheless, it is clear that it can produce some variations in the results, so our suggestion is to use several of them to obtain more robust results. Our work shows that CVIs appear to be better adapted to the VI and Jaccard partition similarity measures than to Adjusted Rand.

An statistical significance analysis of the results showed that there are three main groups of indices and the indices in the first group — Silhouette, Davies–Bouldin, Calinski–Harabasz, generalized Dunn, COP and SDbw — behave better than indices in the last group — Dunn and its Point Symmetry-Distance based variation, Gamma, C-Index, Negentropy increment and OS-Index — being the differences statistically significant.

This work also raises some questions and, therefore, suggests some future work. It is obvious that this type of work can always be improved. Although we consider that we performed an extensive comparison there is room for extending it to include more CVIs, datasets, clustering algorithms and so on. In this context noise and overlap would appear to be the most interesting factors to analyse in greater depth. We also limited this work to crisp clustering, so a fuzzy CVI comparison would be a natural continuation. The analysis of some other kind of indices, such as stability based ones, would also be of great interest.

Finally, we argued that statistical tests are a very valuable tool in data mining and that an effort should be made to use them more widely in clustering. We adapted a method widely accepted in the supervised learning area for our work, but this is just a first approach to the problem and there is a vast field of theoretical research to be addressed.

Acknowledgements

This work was funded by the University of the Basque Country, general funding for research groups (Aldapa, GIU10/02), by the Science and Education Department of the Spanish Government (ModelAccess project, TIN2010-15549), by the Basque Government's SAIOTEK program (Datacc project, S-PE11UN097) and by the Diputación Foral de Gipuzkoa (Zer4you project, DG10/5).

References

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

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

[3] B. Mirkin, Clustering for Data Mining: A Data Recovery Approach, Chapman & Hall/CRC, Boca Raton, Florida, 2005.

[4] P.H.A. Sneath, R.R. Sokal, Numerical Taxonomy, Books in Biology, W.H. Freeman and Company, San Francisco, 1973.

[5] K.J. Holzinger, H.H. Harman, Factor Analysis, University of Chicago Press, Chicago, 1941.

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

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

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

[9] I. Guyon, U. von Luxburg, R.C. Williamson, Clustering: science or art?, in: NIPS 2009 Workshop on Clustering Theory, Vancouver, Canada, 2009.

[10] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, E.R. Dougherty, Model-based evaluation of clustering validation measures, Pattern Recognition 40 (2007) 807–824.

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

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

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

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

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

[16] A. Ben-Hur, A. Elisseeff, I. Guyon, A stability based method for discovering structure in clustered data, in: Biocomputing 2002 Proceedings of the Pacific Symposium, vol. 7, 2002, pp. 6–17.

[17] A.K. Jain, J. Moreau, Bootstrap technique in cluster analysis, Pattern Recognition 20 (1987) 547–568.

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

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

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

[21] M. Halkidi, M. Vazirgiannis, A density-based cluster validity approach using multi-representatives, Pattern Recognition Letters 20 (2008) 773–786.

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

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

[24] K.R. Žalik, B. Žalik, Validity index for clusters of different sizes and densities, Pattern Recognition Letters 32 (2011) 221–234.

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

[26] S. Saha, S. Bandyopadhyay, Performance evaluation of some symmetry-based cluster validity indexes, IEEE Transactions on Systems, Man, and Cybernetics, Part C 39 (2009) 420–425.

[27] R.C. Dubes, How many clusters are best? – an experiment, Pattern Recognition 20 (1987) 645–663.

[28] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, J.M. Pérez, J.I. Martín, Towards a standard methodology to evaluate internal cluster validity indices, Pattern Recognition Letters 32 (2011) 505–515.

[29] J.C. Bezdek, W.Q. Li, Y. Attikiouzel, M. Windham, A geometric approach to cluster validity for normal mixtures, Soft Computing—A Fusion of Foundations, Methodologies and Applications 1 (1997) 166–179.

[30] S. Bandyopadhyay, S. Saha, A point symmetry-based clustering technique for automatic evolution of clusters, IEEE Transactions on Knowledge and Data Engineering 20 (2008) 1441–1457.

[31] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1985) 193–218.

[32] D.-J. Kim, Y.-W. Park, D.-J. Park, A novel validity index for determination of the optimal number of clusters, IEICE Transactions on Information and Systems E84-D (2001) 281–285.

[33] C.A. Sugar, G.M. James, Finding the number of clusters in a dataset, Journal of the American Statistical Association 98 (2003) 750–763.

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

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

[36] P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65.

[37] J.C. Bezdek, N.R. Pal, Some new indexes of cluster validity, IEEE Transactions on Systems, Man, and Cybernetics, Part B 28 (1998) 301–315.

[38] M. Halkidi, M. Vazirgiannis, Clustering validity assessment: finding the optimal partitioning of a data set, in: Proceedings of the First IEEE International Conference on Data Mining (ICDM'01), California, USA, 2001, pp. 187–194.

[39] S. Saitta, B. Raphael, I. Smith, A bounded index for cluster validity, in: P. Perner (Ed.), Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science, vol. 4571, Springer, Berlin, Heidelberg, 2007, pp. 174–187.

[40] I. Gurrutxaga, I. Albisua, O. Arbelaitz, J.I. Martín, J. Muguerza, J.M. Pérez, I. Perona, SEP/COP: an efficient method to find the best partition in hierarchical clustering based on a new cluster validity index, Pattern Recognition 43 (2010) 3364–3373.

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

[42] M. Meilă, Comparing clusterings by the variation of information, in: Proceedings of the Sixteenth Annual Conference on Computational Learning Theory (COLT), 2003, pp. 173–187.

[43] A. Frank, A. Asuncion, UCI machine learning repository, 2010.

[44] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.

[45] T.G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation 10 (1998) 1895–1924.

[46] S. García, F. Herrera, An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons, Journal of Machine Learning Research 9 (2008) 2677–2694.

Olatz Arbelaitz received the M.Sc. and Ph.D. degrees in Computer Science from the University of the Basque Country in 1993 and 2002, respectively. She is an Associate Professor in the Computer Architecture and Technology Department of the University of the Basque Country. She has worked in autonomous robotics, combinatorial optimization and supervised and unsupervised machine learning techniques, focusing lately in web mining.

Ibai Gurrutxaga received the M.Sc. and Ph.D. degrees in Computer Science from the University of the Basque Country in 2002 and 2010. He is an Associate Professor in the Computer Architecture and Technology Department of the University of the Basque Country. He is working in data mining and pattern recognition, focusing on supervised and unsupervised classification (decision trees, clustering, computer security and intrusion detection), and high performance computing.

Javier Muguerza received the M.Sc. and Ph.D. degrees in Computer Science from the University of the Basque Country in 1990 and 1996, respectively. He is an Associate Professor in the Computer Architecture and Technology Department of the University of the Basque Country. His research interests include data mining, pattern recognition and high performance computing.

Jesús María Pérez received the M.Sc. and Ph.D. degrees in Computer Science from the University of the Basque Country in 1993 and 2006, respectively. He is an Associate Professor in the Computer Architecture and Technology Department of the University of the Basque Country. His research interests include data mining and pattern recognition techniques, focusing on classifiers with explanation capacities, learning from imbalanced data and statistical analysis.

Iñigo Perona received the M.Sc. degree in Computer Science from the University of the Basque Country in 2008. He is granted to pursue the Ph.D. at the Computer Architecture and Technology Department of the University of the Basque Country. He is working in data mining and pattern recognition, focusing on supervised and unsupervised classification (web mining, clustering, computer security and intrusion detection).