Relative Clustering Validity Criteria: A Comparative Overview

Lucas Vendramin, Ricardo J. G. B. Campello* and Eduardo R. Hruschka

Department of Computer Sciences of the University of São Paulo at São Carlos, C.P. 668, São Carlos, Brazil

Received 23 October 2009; revised 30 April 2010; accepted 8 May 2010
DOI:10.1002/sam.10080
Published online 30 June 2010 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Many different relative clustering validity criteria exist that are very useful in practice as quantitative measures for evaluating the quality of data partitions, and new criteria have still been proposed from time to time. These criteria are endowed with particular features that may make each of them able to outperform others in specific classes of problems. In addition, they may have completely different computational requirements. Then, it is a hard task for the user to choose a specific criterion when he or she faces such a variety of possibilities. For this reason, a relevant issue within the field of clustering analysis consists of comparing the performances of existing validity criteria and, eventually, that of a new criterion to be proposed. In spite of this, the comparison paradigm traditionally adopted in the literature is subject to some conceptual limitations. The present paper describes an alternative, possibly complementary methodology for comparing clustering validity criteria and uses it to make an extensive comparison of the performances of 40 criteria over a collection of 962,928 partitions derived from five well-known clustering algorithms and 1080 different data sets of a given class of interest. A detailed review of the relative criteria under investigation is also provided that includes an original comparative asymptotic analysis of their computational complexities. This work is intended to be a complement of the classic study reported in 1985 by Milligan and Cooper as well as a thorough extension of a preliminary paper by the authors themselves. © 2010 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 3: 209–235, 2010

Keywords: clustering; validation; relative criteria

1. INTRODUCTION

Data clustering is a fundamental conceptual problem in data mining, in which one aims at determining a finite set of categories to describe a data set according to similarities among its objects [1–3]. The solution to this problem often constitutes the final goal of the mining procedure—having broad applicability in areas that range from image and market segmentation to document categorization and bioinformatics (e.g. see refs [2,4,5])—but solving a clustering problem may also help solving other related problems, such as pattern classification and rule extraction from data [6].

Clustering techniques can be broadly divided into three main types [7]: overlapping, partitional, and hierarchical. The last two are related to each other in that a hierarchical clustering is a nested sequence of hard partitional clusterings, each of which represents a partition of the data set into a different number of mutually disjoint subsets. A hard partition of a data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, composed of n-dimensional feature or attribute vectors $\mathbf{x}_j$, is a collection $\mathbf{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ of k nonoverlapping data subsets $\mathbf{C}_i$ (clusters) such that $\mathbf{C}_1 \cup \mathbf{C}_2 \cup \ldots \cup \mathbf{C}_k = \mathbf{X}$, $\mathbf{C}_i \neq \oslash$, and $\mathbf{C}_i \cap \mathbf{C}_l = \oslash$ for $i \neq l$. Overlapping techniques search for soft or fuzzy partitions by somehow relaxing the mutual disjointness constraints $\mathbf{C}_i \cap \mathbf{C}_l = \oslash$.

The literature on data clustering is extensive. Several clustering algorithms with different characteristics and for different purposes have been proposed and investigated over the past four decades or so [4,5,8,9]. Despite the outstanding evolution of this area and all the achievements obtained during this period of time, a critical issue that is still on the agenda regarding the estimation of the number of clusters contained in data. Most of the clustering algorithms, in particular the most traditional and popular ones, require that the number of clusters be defined either *a priori* or *a posteriori* by the user. Examples are the well-known k-means [7,10], EM (expectation maximization) [11,12], and hierarchical clustering algorithms [2,7]. This is quite restrictive in practice since the number of clusters is generally unknown, especially for n-dimensional data, where visual inspection is prohibitive for "large" n [2,7]. A widely

Additional Supporting Information may be found in the online version of this article.

Correspondence to: Ricardo J. G. B. Campello (campello@icmc.usp.br) known and simple approach to get around this drawback consists of getting a set of data partitions with different numbers of clusters and then selecting that particular partition that provides the best result according to a specific quality criterion [7]. Such a set of partitions may result directly from a hierarchical clustering dendrogram or, alternatively, from multiple runs of a partitional algorithm (e.g. *k*-means) starting from different numbers and initial positions of cluster prototypes.

Many different clustering validity measures exist that are very useful in practice as quantitative criteria for evaluating the quality of data partitions—e.g. see refs [13,14] and references therein. Some of the most well-known validity measures, also referred to as relative validity (or quality) criteria, are possibly the Davies–Bouldin index [7,15], the variance ratio criterion—VRC (so-called Calinski–Harabasz index) [2,16], Dunn's index [17,18], and the silhouette width criterion (SWC) [1,2,19], just to mention a few. It is a hard task for the user, however, to choose a specific measure when he or she faces such a variety of possibilities. To make things even worse, new measures have still been proposed from time to time. For this reason, a problem that has been of interest over more than two decades consists of comparing the performances of existing clustering validity measures and, eventually, that of a new measure to be proposed. Indeed, various researchers have undertaken the task of comparing performances of clustering validity measures since the 1980s. A cornerstone in this area is the work by Milligan and Cooper [14], who compared 30 different measures through an extensive set of experiments involving several labeled data sets. Twenty four years later, that seminal, outstanding work is still used and cited by many authors who deal with clustering validity criteria. In spite of this, the comparison methodology adopted by Milligan and Cooper is subject to three conceptual problems. First, it relies on the assumption that the accuracy of a criterion can be quantified by the number of times it indicates as the best partition (among a set of candidates), a partition with the *right* number of clusters for a specific data set, over many different data sets for which such a number is known in advance—e.g. by visual inspection of 2D/3D data or by labeling synthetically generated data. This assumption has also been implicitly made by other authors who worked on more recent papers involving the comparison of clustering validity criteria (e.g. see refs [13,20–22]). Note, however, that there may exist numerous partitions of a data set into the *right number* of clusters, but clusters that are very *unnatural* with respect to the spatial distribution of the data. As an example, let us consider a partition of the Ruspini data set [1,23] into four clusters, as shown in Fig. 1. Even though visual inspection suggests that four is an adequate estimate of the number of natural clusters for the Ruspini data, common sense says that the clusters displayed in Fig. 1 are far away from those expected natural clusters. On the other hand, there may exist numerous partitions of a data set into the *wrong number* of clusters, but clusters that exhibit a high degree of compatibility with the spatial distribution of the data—e.g. the natural clusters except for some outliers separated apart as independent small clusters or singletons. An example using the Ruspini data is shown in Fig. 2, where five visually acceptable clusters are displayed.

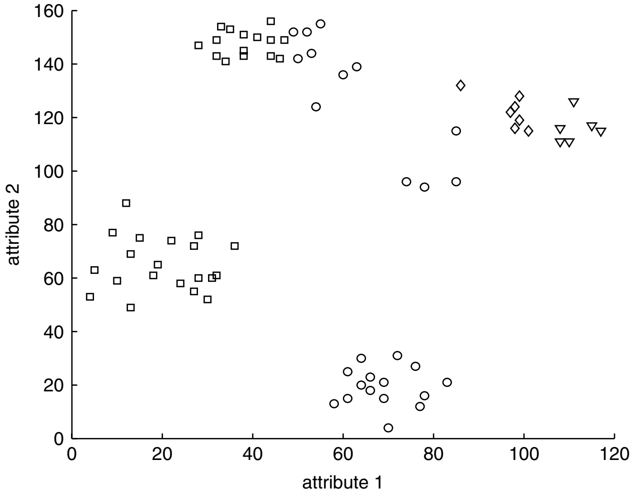


Fig. 1 Unnatural partitioning of the Ruspini data set into four clusters: circles, diamonds, triangles, and squares.

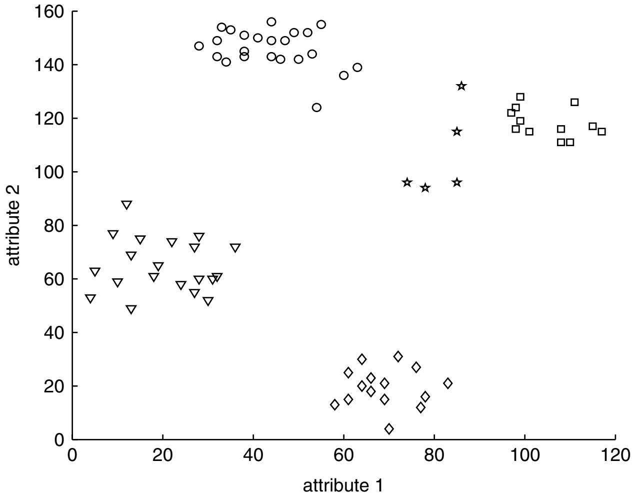


Fig. 2 Partitioning of the Ruspini data into five clusters.

It is important to notice that Milligan and Cooper [14] minimized or even avoided the above-mentioned problem by generating the collection of candidate partitions for each data set in such a way that the *optimal* (known) partition was very likely to be included into that collection as the only partition with the right number of clusters.

This was accomplished by applying a hierarchical clustering algorithm to data sets with well-behaved cluster structures. In spite of this, another potential problem with the methodology adopted in Ref. [14] is that it relies on the assumption that a mistake made by a certain validity criterion when assessing a collection of candidate partitions of a data set can be quantified by the absolute difference between the right (known) number of clusters in the data and the number of clusters contained in the partition elected as the best one. For instance, if a certain criterion suggests that the best partition of a data set with six clusters has eight clusters, then this mistake counts two units for the amount of mistakes made by that specific criterion. However, a partition of a data set with k clusters into $k + \Delta_k$ clusters may be better than another partition of the same data into $k - \Delta_k$ clusters and vice versa. As an example, let us consider a partition of the Ruspini data set into three clusters, as shown in Fig. 3. Such a partition is possibly less visually appealing than that with five clusters shown in Fig. 2 for most people. But more importantly, the partition in Fig. 3 can be deemed worse than that in Fig. 2 from the conceptual viewpoint of cluster analysis. Indeed, the partition in Fig. 3 loses information on the intrinsic structure contained in the data by merging two well-defined clusters, whereas that in Fig. 2 just suggests that some objects nearby a given well-defined cluster would be better as a cluster on their own.

A third potential problem with the methodology adopted in Ref. [14] is that the assessment of each validity criterion relies solely on the correctness (with respect to the number of clusters) of the partition elected as the best one according to that criterion. The accuracy of the criterion when evaluating all the other candidate partitions is just ignored. Accordingly, its capability to properly distinguish among a set of partitions that are not good in general is not taken into account. This capability indicates a particular kind of robustness of the criterion that is important in real-world application scenarios for which no clustering algorithm can provide precise solutions (i.e. compact and separated clusters) due to noise contamination, cluster overlapping, high dimensionality, and other possible complicative factors. Such a robustness is also particularly desirable, for instance, in algorithms for clustering, where an initial random population of partitions evolves toward selectively improving a given clustering validity measure—e.g. see ref. [24] and references therein.

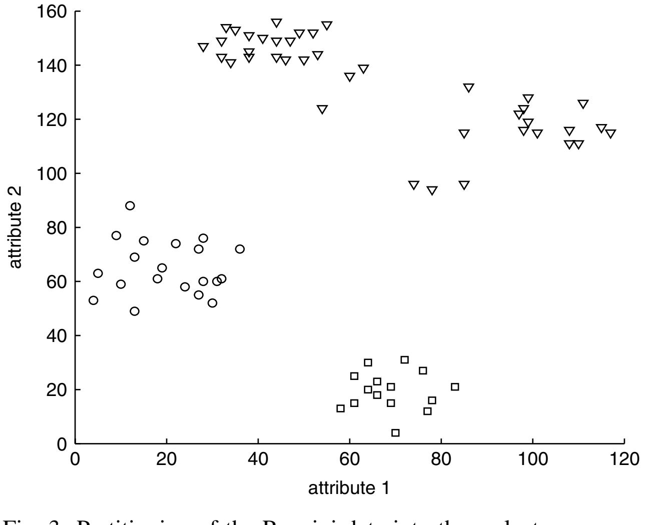


Fig. 3 Partitioning of the Ruspini data into three clusters.

Before proceeding with an attempt at getting around the drawbacks described above, it is important to remark that some very particular criteria are only able to estimate the number of clusters in data—by suggesting when a given iterative (e.g. hierarchical) clustering procedure should stop increasing this number. Such criteria are referred to as stopping rules [14] and should not be seen as clustering validity measures in a broad sense, since they are not able to quantitatively measure the quality of data partitions. In other words, they are not optimization-like validity measures. When comparing the efficacy of stopping rules, the traditional methodology adopted by Milligan and Cooper [14] is possibly the only choice[1]. However, when dealing with optimization-like measures, which are the kind of criterion subsumed in the present paper and can be used as stopping rules as well, a comparison methodology that is broader in scope may also be desired.

The present paper describes an alternative, possibly complementary methodology for comparing relative clustering validity measures and uses it as a basis to perform an extensive comparison study of 40 measures. Given a set of partitions of a particular data set and their respective validity values according to different relative measures, the performance of each measure on the assessment of the whole set of partitions can be evaluated with respect to an external (absolute rather than relative) criterion that supervisedly quantifies the degree of compatibility between each partition and the *right* one, formed by known clusters. Among the most well-known external criteria are the adjusted Rand index (ARI) [7,25] and the Jaccard coefficient [1,7]. It is expected that a good relative clustering validity measure will rank the partitions according to an ordering that is similar to that established by an external criterion, since external criteria rely on supervised information about the underlying structure in the data (known referential clusters). The agreement level between the dispositions of the partitions according to their relative and external validity values can be readily computed using a correlation index—e.g. Pearson coefficient [26]—as formerly envisioned by Milligan

[1] For this reason, stopping rules have not been included into the present study.

in 1981 [27]. Clearly, the larger the correlation value the higher the capability of a relative measure to (unsupervisedly) mirror the behavior of the external index and properly distinguish between better and worse partitions. In particular, if the set of partitions is not good in general (e.g. due to noise contamination), higher correlation values suggest more robustness of the corresponding relative measures.

This work is a thorough extension of a preliminary conference paper [28]. The extensions take place in the following different aspects: (i) 16 additional relative validity measures have been included into the analyses, thus summing up to a total of 40; (ii) a detailed review of these measures have been provided that includes an original comparative asymptotic analysis of their computational complexities; (iii) five different clustering algorithms (four hierarchical ones and k-means) have been used in the experiments, instead of k-means only; (iv) a much more extensive collection of experiments, involving a number of additional data sets (over three times more) and original evaluation scenarios, has been carried out. For instance, the behaviors of the validity criteria when assessing partitions of data sets with different amounts of clusters and dimensions (attributes) have been investigated; (v) part of the study has been devoted to reproducing experiments performed by Milligan and Cooper in their 1985 paper [14], with the inclusion of several measures that were not covered in that reference; and (vi) a discussion on how to transform difference-like validity measures into optimization-like measures has been included.

The remainder of this paper is organized as follows. In Section 2, a collection of relative clustering validity measures are reviewed and their computational complexities are analyzed. Next, in Section 3, a brief review of external clustering validity criteria is presented. In Section 4, an alternative methodology for comparing relative validity measures is described. Such a methodology is then used in Sections 5 and 6 to perform an extensive comparison study involving those measures reviewed in Section 2. Finally, the conclusions and some directions for future research are addressed in Section 7.

2. RELATIVE CLUSTERING VALIDITY CRITERIA

A collection of relative clustering validity criteria from the literature is surveyed in this section, which has been divided into two main subsections. Section 2.1 comprises optimization-like criteria, which are those for which higher (maximization) or lower (minimization) values naturally indicate the best partitions. Section 2.2 addresses difference-like criteria, which are those primarily designed to assess the relative improvement between two consecutive partitions produced by a hierarchical clustering algorithm. Unlike simple stopping rules, which can only be used to halt this sort of algorithm [14], difference-like criteria can be modified so as to exhibit (maximum or minimum) peaks at the partitions deemed the best ones, as will be discussed in Section 2.2.

The description of each criterion in Sections 2.1 and 2.2 is followed by an asymptotic time complexity analysis. Such analysis is based on the assumption that the criterion receives as input a collection of N data objects, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, partitioned into k disjoint clusters, $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k$, each of which is composed of N_l objects ($N_l = cardinality(\mathbf{C}_l) \neq 0$), in such a way that $N_1 + N_2 + \ldots + N_k = N$. It is also assumed that the distance $d(\mathbf{x}_i, \mathbf{x}_j)$ between two data objects can be computed in linear time with respect to the number of dimensions (attributes) of these objects, i.e. it is assumed that $d(\mathbf{x}_i, \mathbf{x}_j)$ is $O(n)$.

Before proceeding with the review of relative clustering validity criteria, it is important to remark that the definitions of most of these criteria subsume the use of numerical data, which means that they operate on data objects described by numerical attributes only ($\mathbf{x}_j = [x_{j1}, \ldots, x_{jn}]^T \in \mathbb{R}^n$). In these cases, the concepts of cluster and/or data centroid may take place. In order to standardize notation, the centroid of a cluster $\mathbf{C}_l$ is hereafter referred to as $\overline{\mathbf{x}}_l$, whereas the centroid of the whole data set (grand mean) is referred to as $\overline{\mathbf{x}}$. So, one has $\overline{\mathbf{x}}_l = [\overline{x}_{l1} \ldots \overline{x}_{ln}]^T = \frac{1}{N_l}\sum_{\mathbf{x}_i \in \mathbf{C}_l} \mathbf{x}_i$ and $\overline{\mathbf{x}} = [\overline{x}_1 \ldots \overline{x}_n]^T = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$. The cost of computing such centroids is included into the complexity analyses of the corresponding criteria.

2.1. Optimization-like Criteria

2.1.1. Calinski–Harabasz (VRC)

The variance ratio criterion [16] evaluates the quality of a data partition as:

$$VRC = \frac{\text{trace}(\mathbf{B})}{\text{trace}(\mathbf{W})} \times \frac{N-k}{k-1} \tag{1}$$

where $\mathbf{W}$ and $\mathbf{B}$ are the $n \times n$ within-group and between-group dispersion matrices, respectively, defined as:

$$\mathbf{W} = \sum_{l=1}^{k} \mathbf{W}_l \tag{2}$$

$$\mathbf{W}_l = \sum_{\mathbf{x}_i \in \mathbf{C}_l} (\mathbf{x}_i - \overline{\mathbf{x}}_l)(\mathbf{x}_i - \overline{\mathbf{x}}_l)^T \tag{3}$$

$$\mathbf{B} = \sum_{l=1}^{k} N_l(\overline{\mathbf{x}}_l - \overline{\mathbf{x}})(\overline{\mathbf{x}}_l - \overline{\mathbf{x}})^T \tag{4}$$

where N_l is the number of objects assigned to the lth cluster, $\overline{\mathbf{x}}_l$ is the n-dimensional vector of sample means within that cluster (cluster centroid), and $\overline{\mathbf{x}}$ is the n-dimensional vector of overall sample means (data centroid). As such, the within-group and between-group dispersion matrices sum up to the scatter matrix of the data set, i.e. $\mathbf{T} = \mathbf{W} + \mathbf{B}$, where $\mathbf{T} = \sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$. The trace of matrix $\mathbf{W}$ is the sum of the within-cluster variances (its diagonal elements). Analogously, the trace of $\mathbf{B}$ is the sum of the between-cluster variances. As a consequence, compact and separated clusters are expected to have small values of trace($\mathbf{W}$) and large values of trace($\mathbf{B}$). Hence, the better the data partition the greater the value of the ratio between trace($\mathbf{B}$) and trace($\mathbf{W}$). The normalization term $(N - k)/(k - 1)$ prevents this ratio to increase monotonically with the number of clusters, thus making VRC an optimization (maximization) criterion with respect to k.

It is worth remarking that it is not necessary to perform the computationally intensive calculation of $\mathbf{W}$ and $\mathbf{B}$ in order to get their traces. Actually, these traces can be readily computed as:

$$\text{trace}(\mathbf{W}) = \sum_{l=1}^{k} \text{trace}(\mathbf{W}_l) \tag{5}$$

$$\text{trace}(\mathbf{W}_l) = \sum_{p=1}^{n} \sum_{\mathbf{x}_i \in \mathbf{C}_l} (x_{ip} - \overline{x}_{lp})^2 \tag{6}$$

$$\text{trace}(\mathbf{B}) = \text{trace}(\mathbf{T}) - \text{trace}(\mathbf{W}) \tag{7}$$

$$\text{trace}(\mathbf{T}) = \sum_{p=1}^{n} \sum_{i=1}^{N} (x_{ip} - \overline{x}_p)^2 \tag{8}$$

where x_{ip} is the pth element (attribute) of the ith data object, $\overline{x}_{lp}$ is the pth element of the centroid of the lth cluster, and $\overline{x}_p$ is the pth element of the data centroid.

Complexity analysis: Computing the centroids takes $O(nN)$ time. Computing both trace($\mathbf{T}$) and trace($\mathbf{W}$) also takes $O(nN)$ time (the latter follows from the observation that $nN_1 + \ldots + nN_k = nN$). Hence, it is straightforward to conclude that the overall time complexity for computing the Calinski–Harabasz index in Eq. (1) using Eqs. (5) through (8) is $O(nN)$.

2.1.2. Davies–Bouldin

The Davies–Bouldin index [15] is somewhat related to VRC in that it is also based on a ratio involving within-group and between-group distances. Specifically, the index evaluates the quality of a given data partition as follows:

$$DB = \frac{1}{k} \sum_{l=1}^{k} D_l \tag{9}$$

where $D_l = max_{l \neq m}\{D_{l,m}\}$. Term $D_{l,m}$ is the within-to-between cluster spread for the lth and mth clusters, i.e. $D_{l,m} = (\overline{d}_l + \overline{d}_m)/d_{l,m}$, where $\overline{d}_l$ and $\overline{d}_m$ are the average within-group distances for the lth and mth clusters, respectively, and $d_{l,m}$ is the inter-group distance between these clusters. These distances are defined as $\overline{d}_l = (1/N_l) \sum_{\mathbf{x}_i \in \mathbf{C}_l} ||\mathbf{x}_i - \overline{\mathbf{x}}_l||$ and $d_{l,m} = ||\overline{\mathbf{x}}_l - \overline{\mathbf{x}}_m||$, where $\|\cdot\|$ is a norm (e.g. Euclidean).

Term D_l represents the worst case within-to-between cluster spread involving the lth cluster. Minimizing D_l for all clusters clearly minimizes the Davies–Bouldin index. Hence, good partitions, composed of compact and separated clusters, are distinguished by small values of DB in Eq. (9).

Complexity analysis: Computing the centroids takes $O(nN)$ time. Given the centroids, computing each term $\overline{d}_l$ is $O(nN_l)$ and, for $l = 1, \ldots, k$, $O(nN_1 + \ldots + nN_k) \rightarrow O(nN)$ operations are required. Also, computing each term $d_{l,m}$ is $O(n)$ and, for $l, m = 1, \ldots, k$, $O(nk^2)$ operations are required. Once all terms $d_{l,m}$ have been computed, computing each term D_l is $O(k)$ and, for $l = 1, \ldots, k$, $O(k^2)$ operations are required. Hence, the complexity of DB in Eq. (9) can be written as $O(nN + nk^2 + k^2) \rightarrow O(n(N + k^2))$. If $k^2 << N$, one gets $O(nN)$. If $k \approx N$, however, one gets $O(nN^2)$.

2.1.3. Dunn

Dunn's index [17] is another validity criterion that is also based on geometrical measures of cluster compactness and separation. It is defined as:

$$DN = \min_{\substack{p,q \in \{1,\ldots,k\} \\ p \neq q}} \left\{ \frac{\delta_{p,q}}{\max\limits_{l \in \{1,\ldots,k\}} \Delta_l} \right\} \tag{10}$$

where Δ_l is the *diameter* of the lth cluster and $\delta_{p,q}$ is the *set distance* between clusters p and q. The set distance $\delta_{p,q}$ was originally defined as the minimum distance between a pair of objects across clusters p and q, i.e. $\min_{\mathbf{x}_i \in \mathbf{C}_p}\{\min_{\mathbf{x}_j \in \mathbf{C}_q} ||\mathbf{x}_i - \mathbf{x}_j||\}$, whereas the diameter Δ_l of a given cluster l was originally defined as the maximum distance between a pair of objects within that cluster, i.e. $\max_{\mathbf{x}_i \in \mathbf{C}_l}\{\max_{\mathbf{x}_j \in \mathbf{C}_l} ||\mathbf{x}_i - \mathbf{x}_j||\}$. Note that the definitions of Δ_l and $\delta_{p,q}$ are directly related to the concepts of within-group and between-group distances, respectively. Bearing this in mind, it is straightforward to verify that partitions composed of compact and separated clusters are distinguished by large values of DN in Eq. (10).

Complexity analysis: All we need to compute DN is the distance between every pair of objects in the data set. If both objects belong to the same cluster, the corresponding distance is used to compute Δ_l, otherwise it is used to compute $\delta_{p,q}$. Given that there are $N(N-1)/2$ pairs of objects in a data set with N objects, computing all Δ_l and $\delta_{p,q}$ requires $O(nN^2)$ time. Equation (10) on its own is $O(k^2)$. Then, it is straightforward to conclude that the complexity of Dunn's index in Eq. (10) is $O(nN^2 + k^2)$ and, provided that $k \in \{2, \ldots, N\}$, this complexity reduces to $O(nN^2)$.

2.1.4. *Seventeen variants of Dunn's index*

The original definitions of *set distance* and *diameter* in Eq. (10) were generalized in Ref. [21] giving rise to 17 variants of the original Dunn's index. These variants can be obtained by combining one out of six possible definitions of $\delta_{p,q}$ (the original one plus five alternative definitions) with one out of three possible definitions of Δ_l (the original one plus two alternative definitions). The alternative definitions for the set distance between the pth and qth clusters are:

$$\delta_{p,q} = \max_{\mathbf{x}_i \in \mathbf{C}_p, \mathbf{x}_j \in \mathbf{C}_q} ||\mathbf{x}_i - \mathbf{x}_j|| \tag{11}$$

$$\delta_{p,q} = \frac{1}{N_p N_q} \sum_{\mathbf{x}_i \in \mathbf{C}_p} \sum_{\mathbf{x}_j \in \mathbf{C}_q} ||\mathbf{x}_i - \mathbf{x}_j|| \tag{12}$$

$$\delta_{p,q} = ||\overline{\mathbf{x}}_p - \overline{\mathbf{x}}_q|| \tag{13}$$

$$\delta_{p,q} = \frac{1}{N_p + N_q} \left(\sum_{\mathbf{x}_i \in \mathbf{C}_p} ||\mathbf{x}_i - \overline{\mathbf{x}}_q|| + \sum_{\mathbf{x}_j \in \mathbf{C}_q} ||\mathbf{x}_j - \overline{\mathbf{x}}_p|| \right) \tag{14}$$

$$\delta_{p,q} = \max \left\{ \max_{\mathbf{x}_i \in \mathbf{C}_p} \min_{\mathbf{x}_j \in \mathbf{C}_q} ||\mathbf{x}_i - \mathbf{x}_j||, \max_{\mathbf{x}_j \in \mathbf{C}_q} \min_{\mathbf{x}_i \in \mathbf{C}_p} ||\mathbf{x}_i - \mathbf{x}_j|| \right\} \tag{15}$$

Note that, in contrast to the primary definition of $\delta_{p,q}$ for the original Dunn's index, which is essentially the *single linkage* definition of set distance, expressions (11) and (12) are precisely its *complete linkage* and *average linkage* counterparts. Definition (13), in its turn, is the same as that for the inter-group distance in the Davies–Bouldin index, Eq. (14) is a hybrid of Eqs. (12) and (13), and Eq. (15) is the Hausdorff metric.

The alternative definitions for diameter are:

$$\Delta_l = \frac{1}{N_l(N_l - 1)} \sum_{(\mathbf{x}_i \neq \mathbf{x}_j) \in \mathbf{C}_l} ||\mathbf{x}_i - \mathbf{x}_j|| \tag{16}$$

$$\Delta_l = \frac{2}{N_l} \sum_{\mathbf{x}_i \in \mathbf{C}_l} ||\mathbf{x}_i - \overline{\mathbf{x}}_l|| \tag{17}$$

where Eq. (16) is the average distance among all $N_l(N_l - 1)/2$ pairs of objects of the lth cluster and Eq. (17) is two times the cluster radius, estimated as the average distance among the objects of the lth cluster and its prototype.

Complexity analysis: The time complexities for the variants of Dunn's index can be estimated by combining the individual complexities associated with the respective equations for $\delta_{p,q}$ and Δ_l. By combining these individual complexities and recalling, Eq. (10) on its own is $O(k^2)$, it is possible to show that the overall complexity for computing most of the variants of Dunn's index is the same as that for computing the original index, that is, $O(nN^2)$. The only two exceptions are the variants given by: (i) the combination of Eqs. (13) and (17), which takes $O(nN + nk^2)$ time; and (ii) the combination of Eqs. (14) and (17), which takes $O(nNk)$ time.

2.1.5. *Silhouette width criterion (SWC)*

Another well-known index that is also based on geometrical considerations about compactness and separation of clusters is the SWC [1,19]. In order to define this criterion, let us consider that the jth object of the data set, $\mathbf{x}_j$, belongs to a given cluster $p \in \{1, \ldots, k\}$. Then, let the average distance of this object to all other objects in cluster p be denoted by $a_{p,j}$. Also, let the average distance of this object to all objects in another cluster q, $q \neq p$, be called $d_{q,j}$. Finally, let $b_{p,j}$ be the minimum $d_{q,j}$ computed over $q = 1, \ldots, k$, $q \neq p$, which represents the average dissimilarity of object $\mathbf{x}_j$ to its closest neighboring cluster. Then, the silhouette of the individual object $\mathbf{x}_j$ is defined as:

$$s_{\mathbf{x}_j} = \frac{b_{p,j} - a_{p,j}}{\max\{a_{p,j}, b_{p,j}\}} \tag{18}$$

where the denominator is just a normalization term. The higher $s_{\mathbf{x}_j}$, the better the assignment of $\mathbf{x}_j$ to cluster p. In case p is a singleton, i.e. if it is constituted uniquely by $\mathbf{x}_j$, then it is assumed by convention that $s_{\mathbf{x}_j} = 0$ [1]. This prevents the SWC, defined as the average of $s_{\mathbf{x}_j}$ over $j = 1, 2, \ldots, N$, i.e.

$$SWC = \frac{1}{N} \sum_{j=1}^{N} s_{\mathbf{x}_j} \tag{19}$$

to elect the trivial solution $k = N$ (with each object of the data set forming a cluster on its own) as the best one. Clearly, the best partition is expected to be pointed out when SWC is maximized, which implies minimizing the intra-group distance ($a_{p,j}$) while maximizing the inter-group distance ($b_{p,j}$).

Complexity analysis: All we need to compute SWC is the distance between every pair of objects in the data set. If both objects belong to the same cluster, the corresponding distance is used to compute $a_{p,j}$, otherwise it is used to compute $d_{q,j}$. Given that there are $N(N-1)/2$ pairs of objects in a data set with N objects, computing all $a_{p,j}$ and $d_{q,j}$ requires $O(nN^2)$ time. Once every $d_{q,j}$ is available, computing $b_{p,j}$ for each object is $O(k)$ and, accordingly, $O(Nk)$ operations are required to compute $b_{p,j}$ for all N objects of the data set. Finally, computing Eq. (19) itself is $O(N)$. Then, the computation of the above-mentioned terms altogether take $O(nN^2 + Nk + N) \rightarrow O(nN^2 + Nk)$ time and, as $k \in \{2, \ldots, N\}$, it follows that the overall complexity of the SWC is $O(nN^2)$.

2.1.6. *Alternative silhouette (ASWC)*

A variant of the original silhouette criterion can be obtained by replacing Eq. (18) with the following alternative definition of the silhouette of an individual object [29]:

$$s_{\mathbf{x}_j} = \frac{b_{p,j}}{a_{p,j} + \epsilon} \tag{20}$$

where ϵ is a small constant (e.g. 10^{-6} for normalized data) used to avoid division by zero when $a_{p,j} = 0$. Note that the rationale behind Eq. (20) is the same as that of Eq. (18), in the sense that both are intended to favoring larger values of $b_{p,j}$ and lower values of $a_{p,j}$. The difference lies in the way they do that, linearly in Eq. (18) and nonlinearly in Eq. (20).

Complexity analysis: The overall complexity of the alternative silhouette criterion is precisely the same as that for the original silhouette, that is, $O(nN^2)$.

2.1.7. *Simplified silhouette (SSWC)*

The original silhouette in Eq. (18) depends on the computation of all distances among all objects. Such a computation can be replaced with a simplified one based on distances among objects and cluster centroids. In this case, $a_{p,j}$ in Eq. (18) is redefined as the dissimilarity of the jth object to the centroid of its cluster, p. Similarly, $d_{q,j}$ is computed as the dissimilarity of the jth object to the centroid of cluster q, $q \neq p$, and $b_{p,j}$ becomes the dissimilarity of the jth object to the centroid of its closest neighboring cluster. The idea is to replace average distances with distances to mean points.

Complexity analysis: Computing the centroids takes $O(nN)$ time. Given the centroids, computing $a_{p,j}$ for the jth object demands only the distance of this object to the centroid of its cluster (the pth cluster), which takes $O(n)$ time. Then, $O(nN)$ operations are needed in order to compute $a_{p,j}$ for all N objects of the data set. In addition, computing those $k-1$ terms $d_{q,j}$ associated with the jth object is $O(nk)$, because only the distances of this object to the centroids of clusters $q \neq p$ are required. So, in order to compute these terms for all N objects of the data set requires $O(nkN)$ operations. Terms $b_{p,j}$ can be derived simultaneously to these operations, at a constant additional cost, and computing Eq. (19) itself is $O(N)$. Then, the overall complexity of the simplified silhouette is estimated as $O(nN + nkN + N) \rightarrow O(nkN)$. When $k << N$, as is usual in practical applications, one gets $O(nN)$. Oppositely, if $k \approx N$, then one gets $O(nN^2)$.

2.1.8. *Alternative simplified silhouette (ASSWC)*

An additional variant of the SWC can be derived by combining the alternative and the simplified silhouettes described in Sections 2.1.6 and 2.1.7, respectively, thus resulting in a hybrid of such versions of the original silhouette. Clearly, the complexity remains the same as that for the simplified version, that is, $O(nkN)$.

2.1.9. *PBM*

The criterion known as PBM [30] is also based on the within-group and between-group distances:

$$PBM = \left(\frac{1}{k}\frac{E_1}{E_K}D_K\right)^2 \tag{21}$$

where E_1 denotes the sum of distances between the objects and the grand mean of the data, i.e. $E_1 = \sum_{i=1}^{N} ||\mathbf{x}_i - \overline{\mathbf{x}}||$, $E_K = \sum_{l=1}^{k} \sum_{\mathbf{x}_i \in \mathbf{C}_l} ||\mathbf{x}_i - \overline{\mathbf{x}}_l||$ represents the sum of within-group distances, and $D_K = \max\limits_{l,m=1,\ldots,k} ||\overline{\mathbf{x}}_l - \overline{\mathbf{x}}_m||$ is the maximum distance between group centroids. So, the best partition should be indicated when PBM is maximized, which implies maximizing D_K while minimizing E_K.

Complexity analysis: Computing the centroids and the grand mean point, $\overline{\mathbf{x}}$, is $O(nN)$. The computation of E_1 and E_K is also $O(nN)$, whereas the calculation of D_K is $O(nk^2)$. Thus, the overall computational complexity of PBM is $O(n(N + k^2))$. If $k^2 << N$, then this complexity becomes $O(nN)$. Otherwise, if $k \approx N$, it becomes $O(nN^2)$.

2.1.10. *C-index*

The C-index criterion [31] is based on the within-group distances, as well as on their maximum and minimum possible values:

$$CI = \frac{\theta - \min\theta}{\max\theta - \min\theta} \tag{22}$$

$$\theta = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} q_{i,j} ||\mathbf{x}_i - \mathbf{x}_j|| \tag{23}$$

where $q_{i,j} = 1$ if the ith and jth objects are in the same cluster and $q_{i,j} = 0$ otherwise. The values for $\min\theta$ and $\max\theta$ can be readily obtained from a sorting procedure. In particular, if both the $t = (N(N-1)/2)$ values for $||\mathbf{x}_i - \mathbf{x}_j||$ and t values for $q_{i,j}$ are increasingly sorted, then the summation over the products of their respective elements results in $\max\theta$. As $q_{i,j} \in \{0, 1\}$, this is equivalent to the sum of the w_d greatest values of $||\mathbf{x}_i - \mathbf{x}_j||$, where $w_d = \sum_{l=1}^{k} N_l(N_l - 1)/2$ is the number of whithin-group distances (number of elements $q_{i,j} = 1$). In essence, $\max\theta$ represents the worst case scenario in which any within-group distance in the partition under evaluation would be greater than or equal to any inter-group distance. Oppositely, if the t values for $||\mathbf{x}_i - \mathbf{x}_j||$ are decreasingly sorted, whereas the t values for $q_{i,j}$ are increasingly sorted, then the summation over the products of their respective elements results in $\min\theta$ (best case scenario). Thus, good partitions are expected to have small C-index values.

Complexity analysis: Computing the vector of distances between every pair of objects is $O(nN^2)$, whereas the procedure for sorting it is $O(N^2 log_2 N^2) \rightarrow O(N^2 log_2 N)$. Computing $\max\theta$ ($\min\theta$) as the summation over the $w_d = \sum_{l=1}^{k} N_l(N_l - 1)/2$ first (last) elements of such a vector is $O(\sum_{l=1}^{k} N_l^2)$, which is less computationally demanding than the $O(nN^2)$ operations needed for computing the vector itself. Similarly, computing θ in Eq. (23) is $O(n \sum_{l=1}^{k} N_l^2)$, which is also less computationally demanding than $O(nN^2)$, needed for computing the vector of distances. To summarize, the overall time complexity of this index is $O(nN^2 + N^2 log_2 N) \rightarrow O(N^2(n + log_2 N))$.

2.1.11. Gamma

The criterion known as gamma [27,32] computes the number of *concordant* pairs of objects (S_+), which is the number of times the distance between a pair of objects from the same group is lower than the distance between a pair of objects from different groups, and the number of *discordant* pairs of objects (S_-), which is the number of times the distance between a pair of objects from the same group is greater than the distance between a pair of objects from different groups. By taking into account these quantities, the criterion is defined as:

$$G = \frac{S_+ - S_-}{S_+ + S_-} \tag{24}$$

$$S_+ = \frac{1}{2}\sum_{l=1}^{k} \sum_{\substack{\mathbf{x}_i,\mathbf{x}_j \in \mathbf{C}_l \\ \mathbf{x}_i \neq \mathbf{x}_j}} \frac{1}{2}\sum_{m=1}^{k} \sum_{\substack{\mathbf{x}_p \in \mathbf{C}_m \\ \mathbf{x}_q \notin \mathbf{C}_m}} \delta(||\mathbf{x}_i - \mathbf{x}_j|| < ||\mathbf{x}_p - \mathbf{x}_q||) \tag{25}$$

$$S_- = \frac{1}{2}\sum_{l=1}^{k} \sum_{\substack{\mathbf{x}_i,\mathbf{x}_j \in \mathbf{C}_l \\ \mathbf{x}_i \neq \mathbf{x}_j}} \frac{1}{2}\sum_{m=1}^{k} \sum_{\substack{\mathbf{x}_p \in \mathbf{C}_m \\ \mathbf{x}_q \notin \mathbf{C}_m}} \delta(||\mathbf{x}_i - \mathbf{x}_j|| > ||\mathbf{x}_p - \mathbf{x}_q||) \tag{26}$$

where $\delta(\cdot) = 1$ if the corresponding inequality is satisfied and $\delta(\cdot) = 0$ otherwise. Conceptually speaking, better partitions are expected to have higher values of S_+, lower values of S_-, and, as a consequence, higher values of G in Eq. (24).

Complexity analysis: Since every pair of objects is associated to either a within-group distance or a between-group distance, this criterion needs to compute all pairwise distances between objects, which is $O(nN^2)$. In addition, computing S_+ and S_- demands that all $\sum_{l=1}^{k} N_l(N_l - 1)/2$ within-group distances be compared to all $\sum_{m=1}^{k} N_m(N - N_m)$ between-group distances, resulting in an additional complexity of $O(\sum_{l=1}^{k} N_l(N_l - 1) \cdot \sum_{m=1}^{k} N_m(N - N_m))$. Assuming that the number of objects in each group l is directly proportional to the cardinality of the data set, N, and inversely proportional to the number of groups, k, i.e. $N_l \propto N/k$, this complexity becomes $O(\sum_{l=1}^{k} (\frac{N}{k})^2 \cdot \sum_{m=1}^{k} \frac{N}{k}(N - \frac{N}{k})) \rightarrow O(\frac{N^2}{k} \sum_{m=1}^{k} \frac{N^2(k-1)}{k^2}) \rightarrow O(\frac{N^2}{k} N^2) \rightarrow O(\frac{N^4}{k})$. Thus, the overall complexity for computing gamma is estimated as $O(nN^2 + \frac{N^4}{k})$.

2.1.12. G(+)

G(+) [27,33] is another criterion based on the relationships between *discordant* pairs of objects (S_-). Specifically, it is defined as the proportion of discordant pairs with respect to the maximum number of possible comparisons, $t(t-1)/2$, as follows:

$$G(+) = \frac{2S_-}{t(t-1)} \tag{27}$$

Following the discussions in the previous section, it can be seen that good partitions are expected to have small values for S_- and, as a consequence, small values for G(+) in Eq. (27).

Complexity analysis: The time complexity analysis for G(+) is similar to that performed for gamma in

Section 2.1.11. As such, it also results in an order of magnitude of $O(nN^2 + \frac{N^4}{k})$.

2.1.13. *Tau*

Tau [27,33] is based on the τ correlation [34,35] between the matrix that stores all the distances between pairs of objects and a binary matrix in which each entry indicates whether a given pair of objects belongs to the same cluster (0) or not (1). It is another criterion that can be written in terms of the numbers of *concordant* (S_+) and *discordant* (S_-) pairs of objects, as follows:

$$\tau = \frac{S_+ - S_-}{\sqrt{(t(t-1)/2 - t_{ie})\,(t(t-1)/2)}} \tag{28}$$

where S_+ and S_- are defined in Eqs. (25) and (26), respectively, $t = N(N-1)/2$, and

$$t_{ie} = \binom{w_d}{2} + \binom{b_d}{2} = \frac{w_d(w_d - 1)}{2} + \frac{b_d(b_d - 1)}{2} \tag{29}$$

with $w_d = \sum_{l=1}^{k} N_l(N_l - 1)/2$ and $b_d = \sum_{l=1}^{k} N_l(N - N_l)$ /2. Following the discussions about gamma in Section 2.1.11, it can be noticed that better partitions are presumed to be distinguished by higher values of τ in Eq. (28).

Complexity analysis: The time complexity analysis for tau is similar to that performed for gamma in Section 2.1.11. The final result remains the same, i.e. $O(nN^2 + \frac{N^4}{k})$.

2.1.14. *Point-biserial*

Similarly to tau, point-biserial [14,27] is also based on a correlation measure between a distance matrix and a binary matrix that encodes the mutual memberships of pairs of objects to clusters, as follows:

$$PB = \frac{(d_b - d_w)\sqrt{w_d \cdot b_d / t^2}}{s_d} \tag{30}$$

where d_w is the average intra-group distance, d_b is the average inter-group distance, $t = N(N-1)/2$ is the total number of distances between pairs of objects, s_d is the standard deviation over all those distances, $w_d = \sum_{l=1}^{k} N_l(N_l - 1)/2$ is the number of intra-group distances, and $b_d = \sum_{l=1}^{k} N_l(N - N_l)/2$ is the number of inter-group distances. Conceptually speaking, good partitions should exhibit large inter-group and small intra-group distances. Hence, point-biserial is a maximization criterion.

Complexity analysis: This criterion requires the computation of d_b, d_w, w_d, b_d, t, and s_d. Computing t is $O(1)$. The computation of d_w and d_b, in turn, requires calculating all the distances between pairs of objects. In fact, if a given pair of objects belongs to the same cluster, then the respective distance is used to compute d_w; otherwise, it is employed to calculate d_b. During this procedure, which takes $O(nN^2)$ time, it is also possible to compute w_d and b_d. In summary, computing d_w, d_b, w_d, and b_d is $O(nN^2)$. In order to compute s_d, it is necessary to process $w_d + b_d = t$ distance values, which is also $O(nN^2)$. From these observations, it is straightforward to verify that the time complexity of point-biserial is $O(nN^2)$.

2.1.15. *C/$\sqrt{\text{k}}$*

The criterion known as C/$\sqrt{\text{k}}$ [36,37] is based on the individual contribution of each attribute to the between- and within-cluster variances, as follows:

$$C/\sqrt{\text{k}} = \frac{1}{\sqrt{k}}\frac{1}{n}\sum_{q=1}^{n}\sqrt{\frac{SSB_q}{SST_q}} \tag{31}$$

$$SSB_q = SST_q - SSW_q \tag{32}$$

$$SST_q = \sum_{i=1}^{N} ||x_{iq} - \overline{x}_q||^2 \tag{33}$$

$$SSW_q = \sum_{l=1}^{k}\sum_{\mathbf{x}_i \in \mathbf{C}_l} ||x_{iq} - \overline{x}_{lq}||^2 \tag{34}$$

where x_{iq} is the qth attribute of the ith object, $\overline{x}_q$ the qth attribute of the centroid of the whole data (grand mean, $\overline{\mathbf{x}}$), and $\overline{x}_{lq}$ is the qth attribute of the centroid of the lth cluster, $\overline{\mathbf{x}}_l$. Compact clusters tend to have low within-group variances ($SSW_q, q = 1, \ldots, n$), thus leading to high values of SSB_q/SST_q (close to one). So, higher values of C/$\sqrt{\text{k}}$ are expected to indicate better partitions. Note that, due to the fact that increasing the number of clusters in the partition under evaluation tends to increase the ratio SSB_q/SST_q, the criterion is not only normalized in relation to n, but also in relation to $\sqrt{k}$. This is an attempt at counterbalancing the increase of $\sqrt{SSB_q/SST_q}$ as a function of k. However, increasing k in a persistent way makes $SSW_q \to 0$, $SSB_q \to SST_q$ and, as a consequence, C/$\sqrt{\text{k}}$ $\to$ $1/\sqrt{k}$, which means that this criterion tends to be biased toward disfavoring partitions formed by many clusters, irrespective of their qualities.

Complexity analysis: This criterion requires computing cluster centroids, $\overline{\mathbf{x}}_l$, and the centroid of the whole data, $\overline{\mathbf{x}}$, with a cost of $O(nN)$ operations. In addition, one has to compute the values for SSW_q and SST_q, for $q = 1, \ldots, n$. After computing the centroids, it is possible to calculate SST_q and SSW_q for each individual attribute in $O(N)$ time. Because such quantities must be computed for every attribute, $O(nN)$ operations are required. To summarize, computing C/$\sqrt{\text{k}}$ in Eq. (32) is $O(nN)$.

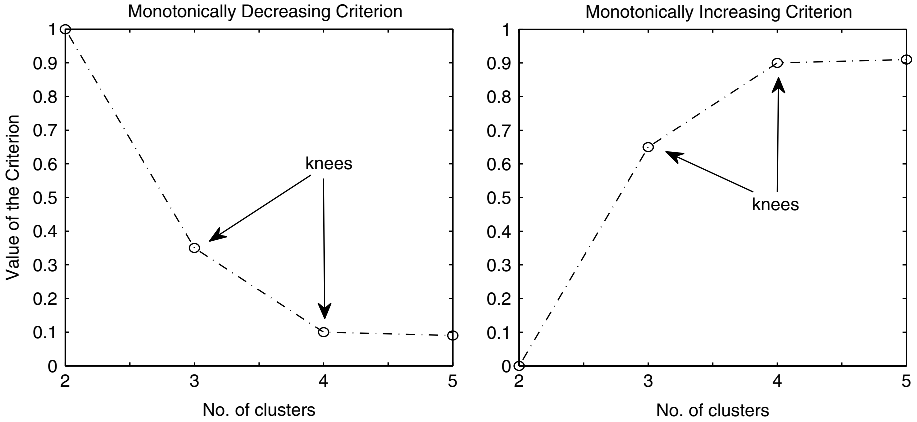


Fig. 4 Examples of criteria with monotonically decreasing (increasing) values as a function of k.

2.2. Difference-like Criteria

Some criteria are used to assess relative improvements on some relevant characteristic of the data (e.g. within-group variances) over a set of successive data partitions produced by a given iterative procedure—usually a hierarchical clustering algorithm. Such criteria can be monotonically increasing (decreasing) as the number of clusters varies from $k = 2$ to $k = N$, as illustrated in Fig. 4. In these cases, one usually tries to identify a prominent "knee" or "elbow" that suggests the most appropriate number of clusters existing in the data. Such criteria are hereafter called *difference-like criteria*, for they require relative assessments between values obtained in two consecutive data partitions (formed by k and $k + 1$ clusters).

Before evaluating the results achieved by *difference-like criteria*, it is necessary to transform them into *optimization-like criteria*, for which an extreme value (peak) reveals the partition elected as the best one among a set of candidates with different values of k. Milligan and Cooper stated in Ref. [14] that they used the *difference between hierarchical levels* to transform *difference-like criteria* into *optimization-like criteria*. However, the authors did not present any formal description for such a concept, for which two possible realizations are:

$$C_{\text{new}}(k) = abs\big(C_{\text{orig}}(k-1) - C_{\text{orig}}(k)\big) \qquad (35)$$

or

$$C_{\text{new}}(k) = abs\Big(abs\big(C_{\text{orig}}(k-1) - C_{\text{orig}}(k)\big) - abs\big(C_{\text{orig}}(k) - C_{\text{orig}}(k+1)\big)\Big) \qquad (36)$$

where $abs(.)$ stands for the absolute value of the argument, $C_{new}(k)$ the value of the transformed criterion for the partition formed by k clusters, and $C_{orig}(k-1)$, $C_{orig}(k)$ and $C_{orig}(k+1)$ are the values of the original difference-like criterion for partitions formed by $k-1$, k, and $k+1$ clusters, respectively. Notice, however, that a "knee" in a chart is recognized by an abrupt *relative* (rather than absolute) change in the *variation* of the value between consecutive partitions. From this observation, we here propose to use the following transformation:

$$C_{\text{new}}(k) = abs\left(\frac{C_{\text{orig}}(k-1) - C_{\text{orig}}(k)}{C_{\text{orig}}(k) - C_{\text{orig}}(k+1)}\right) \qquad (37)$$

Indeed, we will experimentally show in Section 5 that such a transformation provides results significantly better than those found using Eq. (35) or (36). For this reason, the experimental results to be reported in this work are based upon the assumption that *difference-like criteria* are converted into *optimization-like criteria* by Eq. (37).

2.2.1. Trace(W)

Trace(W) is a simple and widely known difference-like criterion, defined as [38]:

$$\mathcal{V}_1 = \text{trace}(\mathbf{W}) \qquad (38)$$

where **W** is the within-group covariance matrix in Eq. (2). ***Complexity analysis:*** From Eqs. (5) and (6) it can be readily verified that trace(W) requires $O(nN)$ operations.

2.2.2. Trace(CovW)

A variant of the trace(W) criterion involves using the *pooled covariance matrix* instead of **W** [14], as follows:

$$\mathcal{V}_2 = \text{trace}(\mathbf{W}_p) \qquad (39)$$

$$\mathbf{W}_p = \frac{1}{N-k}\sum_{l=1}^{k}\mathbf{W}_l \qquad (40)$$

$$\mathbf{W}_l = \sum_{\mathbf{x}_i \in \mathbf{C}_l}(\mathbf{x}_i - \overline{\mathbf{x}}_l)(\mathbf{x}_i - \overline{\mathbf{x}}_l)^T \qquad (41)$$

Complexity analysis: It is straightforward to verify that the asymptotic computational costs of trace(CovW) and trace(W) are the same, namely, $O(nN)$.

2.2.3. Trace($W^{-1}B$)

The criterion known as trace($\mathrm{W}^{-1}\mathrm{B}$) [38] is based both on **W** in Eq. (2) and **B** in Eq. (4), as follows:

$$\mathcal{V}_3 = \mathrm{trace}(\mathbf{W}^{-1}\mathbf{B}) \tag{42}$$

Complexity analysis: Differently from the criteria addressed in Sections 2.2.1 and 2.2.2, this criterion indeed demands the computation of the covariance matrices **W** and **B**, whose time complexities are $O(n^2N)$ and $O(n^2k)$, respectively. Provided that computing the inverse of **W** and its multiplication by **B** can both be accomplished in $O(n^3)$ time, and given that the computation of the trace of the resulting matrix is $O(n)$, it follows that the time complexity of trace($\mathrm{W}^{-1}\mathrm{B}$) is $O(n^2N + n^2k + n^3)$. Since $k \in \{2, \ldots, N\}$, computing trace($\mathrm{W}^{-1}\mathrm{B}$) is then $O(n^2N + n^3)$.

2.2.4. $|T|/|W|$

|T|/|W| [38] is a criterion that uses information from the determinants of the data covariance matrix and within-group covariance matrix, as follows:

$$\mathcal{V}_4 = \frac{|\mathbf{T}|}{|\mathbf{W}|} \tag{43}$$

where $|\cdot|$ stands for the determinant and matrices **W** and $\mathbf{T} = \mathbf{W} + \mathbf{B}$ are the same as previously defined in Section 2.1.1.

Complexity analysis: The computation of |T|/|W| requires calculating the covariance matrices **T** and **W**, whose overall complexity is $O(n^2N)$. Once computing the determinants of these matrices requires $O(n^3)$ operations, the overall complexity of the |T|/|W| criterion is $O(n^2N + n^3)$.

2.2.5. N log($|T|/|W|$)

A variant of the |T|/|W| criterion just described involves using its logarithmic transformation. More precisely, the Nlog(|T|/|W|) criterion [39,40] is given by:

$$\mathcal{V}_5 = N \log_{10}\left(\frac{|\mathbf{T}|}{|\mathbf{W}|}\right) \tag{44}$$

Complexity analysis: Following the analysis described in Section 2.2.4, it can be shown that computing Nlog(|T|/|W|) is $O(n^2N + n^3)$.

2.2.6. $k^2|W|$

The criterion known as $\mathrm{k}^2|W|$ [41] is also based on the determinant of the within-group covariance matrix, as follows:

$$\mathcal{V}_6 = k^2|\mathbf{W}| \tag{45}$$

Complexity analysis: Computing $\mathrm{k}^2|\mathrm{W}|$ requires calculating the within-group covariance matrix, **W**, whose complexity is $O(n^2N)$, and its determinant, which demands $O(n^3)$ operations. Thus, the overall computational complexity of $\mathrm{k}^2|\mathrm{W}|$ in Eq. (45) is $O(n^2N + n^3)$.

2.2.7. log(SSB/SSW)

The log(SSB/SSW) criterion [8] makes use of the within- and between-group distances, as follows:

$$\mathcal{V}_7 = \log_{10}\left(\frac{SSB}{SSW}\right) \tag{46}$$

where $SSW = \sum_{l=1}^{k}\sum_{\mathbf{x}_i \in \mathbf{C}_l} ||\mathbf{x}_i - \bar{\mathbf{x}}_l||^2$ and $SSB = \sum_{l=1}^{k-1}\sum_{m=l+1}^{k} \frac{||\bar{\mathbf{x}}_l - \bar{\mathbf{x}}_m||^2}{(1/N_l)+(1/N_m)}$.

Complexity analysis: Computing the centroids is $O(nN)$. SSW requires computing the distances between every object and the centroid of the cluster it belongs to, which is also $O(nN)$, whereas for SSB it is necessary to calculate every distance between pairs of centroids, leading to a computational cost of $O(nk^2)$. Thus, the overall computational complexity of log(SSB/SSW) is $O(n(k^2 + N))$. If $k^2 << N$, then it is $O(nN)$. If $k \approx N$, however, it becomes $O(nN^2)$.

2.2.8. Ball and Hall

The Ball–Hall criterion [42] is just the well-known within-group sum of distances:

$$\mathcal{V}_8 = \frac{1}{N}\sum_{l=1}^{k}\sum_{\mathbf{x}_i \in \mathbf{C}_l} ||\mathbf{x}_i - \bar{\mathbf{x}}_l|| \tag{47}$$

Complexity analysis: Computing centroids and N within-group distances is $O(nN)$. Thus, Ball–Hall is $O(nN)$.

2.2.9. McClain and Rao

McClain–Rao [43,44] is another criterion based on within- and between-group distances, as follows:

$$\mathcal{V}_9 = \frac{B/\left(N^2 - \sum_{l=1}^{k} N_l^2\right)}{W/\left([\sum_{l=1}^{k} N_l^2] - N\right)} \tag{48}$$

where B and W are given by:

$$B = \sum_{l=1}^{k-1} \sum_{m=l+1}^{k} \sum_{\mathbf{x}_i \in \mathbf{C}_l} \sum_{\mathbf{x}_j \in \mathbf{C}_m} ||\mathbf{x}_i - \mathbf{x}_j|| \tag{49}$$

$$W = \frac{1}{2} \sum_{l=1}^{k} \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{C}_l} ||\mathbf{x}_i - \mathbf{x}_j|| \tag{50}$$

Milligan and Cooper [14] evaluated McClain–Rao as an optimization-like criterion. However, we performed a collection of preliminary experiments in which McClain–Rao performed significantly better (eight times more accurately) when considered as a difference-like criterion, transformed into an optimization-like one. For this reason, McClain–Rao is deemed a difference-like criterion in the experiments to be reported in this work.

Complexity analysis: Computing centroids is $O(nN)$. In addition, it is necessary to calculate all the distances between pairs of objects. In fact, if a given pair of objects belongs to the same cluster, then the respective distance is used to compute W; otherwise, it is employed to calculate B. This procedure demands $O(nN^2)$ operations. Hence, the overall computational complexity of McClain–Rao is $O(nN^2)$.

2.3. Summary of the Complexity Analyses

The computational complexity associated with each relative clustering validity criteria reviewed in the previous sections is displayed in Table 1[2].

Table 1. Computational complexities of several relative validity criteria—difference-like criteria are signed with $*$.

–	Criterion	Complexity
	Calinski–Harabasz (VRC)	$O(nN)$ [Eqs. (5)–(8)]
	Davies–Bouldin (DB)	$O(n(k^2 + N))$
	Dunn	$O(nN^2)$
	Silhouette width criterion (SWC)	$O(nN^2)$
	Alternative silhouette (ASWC)	$O(nN^2)$
	Simplified silhouette (SSWC)	$O(nNk)$
	Alternative simplified silhouette (ASSWC)	$O(nNk)$
	PBM	$O(n(k^2 + N))$
	C-index	$O(N^2(n + log_2 N))$
	Gamma	$O(nN^2 + N^4/k])$
	G(+)	$O(nN^2 + N^4/k])$
	Tau	$O(nN^2 + N^4/k])$
	Point-biserial	$O(nN^2)$
	$\mathrm{C}/\sqrt{\mathrm{k}}$	$O(nN)$
*	Trace(W)	$O(nN)$
*	Trace(CovW)	$O(nN)$
*	Trace($\mathrm{W}^{-1}\mathrm{B}$)	$O(n^2N + n^3)$
*	|T|/|W|	$O(n^2N + n^3)$
*	Nlog(|T|/|W|)	$O(n^2N + n^3)$
*	$\mathrm{k}^2\mathrm{W}$	$O(n^2N + n^3)$
*	log(SSB/SSW)	$O(n(k^2 + N))$
*	Ball–Hall	$O(nN)$
*	McClain–Rao	$O(nN^2)$

3. EXTERNAL CLUSTERING VALIDITY CRITERIA

In this section, external validity criteria that will be used further in this work—as part of the methodology adopted for comparing relative criteria—are also reviewed.

3.1. Rand Index

The Rand index [45] can be seen as an absolute criterion that allows the use of properly labeled data sets for performance assessment of clustering results. This very simple and intuitive index handles two hard partition matrices (R and Q) of the same data set. The reference partition, R, encodes the class labels, i.e. it partitions the data into k^* known categories of objects. Partition Q, in turn, partitions the data into k clusters, and is the one to be evaluated. Given the above remarks, the Rand index is then defined as [2,7,45]:

$$\omega = \frac{a + d}{a + b + c + d} \tag{51}$$

where:

- a: Number of pairs of data objects belonging to the same class in R and to the same cluster in Q.
- b: Number of pairs of data objects belonging to the same class in R and to different clusters in Q.
- c: Number of pairs of data objects belonging to different classes in R and to the same cluster in Q.
- d: Number of pairs of data objects belonging to different classes in R and to different clusters in Q.

Terms a and d are measures of consistent classifications (agreements), whereas terms b and c are measures of inconsistent classifications (disagreements). Note that: (*i*) $\omega \in [0, 1]$; (*ii*) $\omega = 0$ iff Q is completely inconsistent, i.e. $a = d = 0$; and (*iii*) $\omega = 1$ iff the partition under evaluation matches exactly the reference partition, i.e. $b = c = 0$ ($Q \equiv R$).

[2] Please, refer to Section 2.1.4 for the complexities of the variants of Dunn's index.

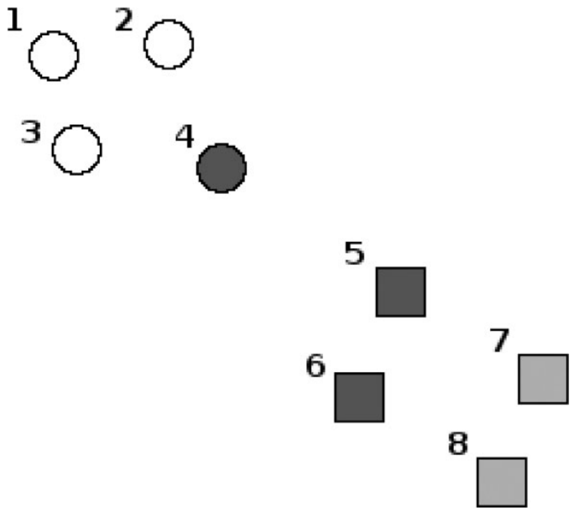


Fig. 5 Data set with $k^* = 2$ classes (class 1 in circles and class 2 in squares) partitioned into $k = 3$ clusters (clusters 1, 2, and 3 in black, white, and gray, respectively).

As an example, consider the data set in Fig. 5. It is composed of two classes with four objects each. Class 1 (in circles) is composed of objects 1, 2, 3, and 4, while class 2 (in squares) is composed of objects 5, 6, 7, and 8. As shown in Fig. 5, the same data set has been partitioned into three clusters—Cluster 1 in black (objects 4, 5, and 6), Cluster 2 in white (objects 1, 2, and 3), and Cluster 3 in gray (objects 7 and 8). The pairs of data objects belonging to the same class and to the same cluster are therefore (1,2), (1,3), (2,3), (5,6), and (7,8). Thus, the number of pairs of objects in the same class and in the same cluster is $a = 5$. Similarly, the other terms in Eq. (51) can easily be computed as $b = 7$, $c = 2$, and $d = 14$, which results in a Rand index of $\omega = 0.6786$.

3.2. Adjusted Rand Index

One of the main criticisms against the original Rand index is that it is not *corrected for chance*, that is, its expected value is not zero when comparing random partitions. Correcting an index for chance means normalizing it so that its (expected) value is 0 when the partitions are selected by chance and 1 when a perfect match is achieved [7]. Hubert and Arabie derived the following ARI [25]:

$$\omega_A = \frac{a - \frac{(a+c)(a+b)}{M}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)(a+b)}{M}} \tag{52}$$

where $M = a + b + c + d$. This index is corrected for chance under the assumption that the number of groups (classes/clusters) in both partitions R and Q to be compared is the same.

3.3. Jaccard Coefficient

Another common criticism against the original Rand index is that it gives the same importance to both the agreement terms a and d, thus making no difference between pairs of objects that are joined or separated in both the referential and evaluated partitions [46]. This policy is arguable, particularly if a partition is interpreted as a set of groups of joined elements, the separations being just consequences of the grouping procedure [47]. This interpretation suggests that term d should be removed from the formulation of the Rand index. Indeed, it is known that this term may dominate the other three terms (a, b, and c), thus causing the Rand index to become unable to properly distinguish between good and bad partitions [48,49]. This situation gets particularly critical as the number of classes/clusters increases, since the value of the index tends to increase too [2,50,51]. The ordinary removal of term d from the original Rand index in Eq. (51) results in the so-called Jaccard coefficient, defined as [1,7]:

$$\omega_J = \frac{a}{a + b + c} \tag{53}$$

Clearly, the rationale behind the Jaccard coefficient in Eq. (53) is essentially the same as that for the Rand index, except for the absence of term d (which does not affect normality, i.e. $\omega_J \in [0, 1]$). An interesting interpretation of the differences between these two indexes arises when d is viewed as a "neutral" term—counting pairs of objects that are not clearly indicative either of similarity or of inconsistency—in contrast to the others, viewed as counts of "good pairs" (term a) and "bad pairs" (terms b and c) [51]. From this viewpoint, the Jaccard coefficient can be seen as a proportion of good pairs with respect to the sum of non-neutral (good plus bad) pairs, whereas the Rand index is just the proportion of pairs not definitely bad with respect to the total number of pairs.

4. COMPARING RELATIVE VALIDITY CRITERIA

As previously discussed in Section 1, the methodology usually adopted in the literature to compare relative clustering validity criteria, based essentially on the ability of the indexes to indicate the *right* number of clusters in a data set, is subject to conceptual limitations. In this work, an alternative, possibly complementary methodology [28] is adopted. Such a methodology can be better explained by means of a pedagogical example. For the sake of simplicity and without any loss of generality, let us assume in this example that our comparison involves only the performances of two validity criteria in the evaluation of a small set of partitions of a single labeled data set. The hypothetical results

Table 2. Example of relative and external evaluation results of five partitions of a data set.

Partition	Relative criterion 1	Relative criterion 2	External criterion
1	0.75	0.92	0.82
2	0.55	0.22	0.49
3	0.20	0.56	0.31
4	0.95	0.63	0.89
5	0.60	0.25	0.67

are displayed in Table 2, which shows the values of the relative criteria under investigation as well as the values of a given external criterion (e.g. Jaccard or ARI described in Section 3) for each of the five data partitions available.

Assuming that the evaluations performed by the external criterion are trustable measures of the quality of the partitions[3], it is expected that the better the relative criterion the greater its capability of evaluating the partitions according to an ordering and magnitude proportions that are similar to those established by the external criterion. Such a degree of similarity can be straightforwardly computed using a sequence correlation index, such as the well-known Pearson product-moment correlation coefficient [26]. Clearly, the larger the correlation value the higher the capability of a relative measure to unsupervisedly mirror the behavior of the external index and properly distinguish between better and worse partitions. In the example of Table 2, the Pearson correlation between the first relative criterion and the external criterion (columns 2 and 4, respectively) is 0.9627. This high value reflects the fact that the first relative criterion ranks the partitions in the same order that the external criterion does. Unitary correlation is not reached only because there are some differences in the relative importance (proportions) given to the partitions. Contrariwise, the correlation between the second relative criterion and the external criterion scores is only 0.4453. This is clearly in accordance with the strong differences that can be visually observed between the evaluations in columns 3 and 4 of Table 2.

In a practical comparison procedure, there should be multiple partitions of varied qualities for a given data set. Moreover, a practical comparison procedure should involve a representative collection of different data sets that fall within a given class of interest (e.g. mixtures of Gaussians). This way, if there are N_D labeled data sets available, then there will be N_D correlation values associated with each relative validity criterion, each of which represents the agreement level between the evaluations of the partitions of one specific data set when performed by that relative criterion and by an external criterion. The mean of such N_D correlations is a measure of resemblance between those particular (relative and external) validity criteria, at least with respect to that specific collection of data sets. Despite this, besides just comparing such means for different relative criteria in order to rank them, one should also apply to the results an appropriate statistical test to check the hypothesis that there are (or aren't) significant differences among those means. In summary the procedure is given below:

1. Take N_D different data sets with known clusters.
2. For each data set, get a collection of N_π data partitions of varied qualities and numbers of clusters. For instance, such partitions can be obtained from a single run of a hierarchical algorithm or from multiple runs of a partitional algorithm with different numbers and initial positions of prototypes.
3. For each data set, compute the values of the relative and external validity criteria for each of the N_π partitions available. Then, for each relative criterion, compute the correlation between the corresponding vector with N_π relative validity values and the vector with N_π external validity values.
4. For each relative criterion, compute the mean[4] of its N_D correlation values (one per data set). Then, rank all the criteria according to their means and apply an appropriate statistical test to the results to check whether the differences between the means are significant or not from a statistical perspective, i.e. when taking variances into account.

Remark 1. Since the external validity criteria are typically maximization criteria, minimization relative criteria, such as Davies–Bouldin, G(+), and C-index, must be converted into maximization ones. To do so, flip the values of such criteria around their means before computing their correlation with the external validity values.

Remark 2. The comparison methodology described above is a generalization of the one proposed in a 1981 paper by Milligan [27], in which the author stated that “Logically, if a given criterion is succeeding in indicating the degree of correct cluster recovery, the index should exhibit a close association with an external criterion which reflects the actual degree of recovery of structure in the proposed partition solution”. Milligan, however, conjectured that the above statement would possibly not be justified for comparisons involving partitions with different numbers of clusters, due to a kind of monotonic trend that some external

[3] This is based on the fact that such a sort of criterion relies on information about the known referential clusters.

[4] Or another statistic of interest.

indexes may exhibit as a function of this quantity. For this reason, the analyses in Ref. [27] were limited to partitions with the right number (k^*) of clusters known to exist in the data. This limitation itself causes two major impacts on the reliability of the comparison procedure. First, the robustness of the criteria under investigation in terms of their ability to properly distinguish among partitions that are not good in general is not taken into account. Second, if the partitions are obtained from a hierarchical algorithm, taking only k^* implies that there will be a single value of a given relative criterion associated with each data set, which means that a single correlation value will result from the pairs of values of relative and external criteria computed over the whole collection of available data sets. Obviously, a single value does not allow statistical evaluations of the results.

The more general methodology adopted here rules out those negative impacts resulting from Milligan's conjecture mentioned above. But what about the conjecture? Such a conjecture is possibly one of the reasons that made Milligan and Cooper not to adopt the same idea in their 1985 paper [14], which was focused on procedures for determining the number of clusters in data. In our opinion, such a conjecture was not properly elaborated. In fact, while it is true that some external indexes (e.g. the original Rand index described in Section 3) do exhibit a monotonic trend as a function of the number of clusters, such a trend is observed when the number of clusters of both the referential and evaluated partitions increase [50,51]—or in specific situations in which there is no structure in the data and in the corresponding referential partition [52]. Neither of these is the case, however, when one takes a well-defined referential partition of a structured data set with a fixed number of clusters and compares it against partitions of the same data produced by some clustering algorithm with variable number of clusters, as verified later by Milligan and Cooper themselves [52].

5. EXPERIMENTAL RESULTS—PART I

The collection of experiments to be described in this work is divided into two parts. The first one (this section) involves the comparison of those 40 relative validity criteria reviewed in Section 2 using precisely the same clustering algorithms and faithful reproductions of the synthetic data sets adopted in the studies by Milligan and Cooper [14,27].

5.1. Data Sets

The data sets adopted here are reproductions of the artificial data sets used in Ref. [14,27]. The data generator was developed following strictly the descriptions in those references. In brief, the data sets consist of a total of $N = 50$ objects each, embedded in either an $n = 4$, 6, or 8 dimensional Euclidean space. Each data set contains either $k^* = 2$, 3, 4, or 5 distinct clusters for which overlap of cluster boundaries is permitted in all but the first dimension of the variables space. The actual distribution of the objects within clusters follows a (mildly truncated) multivariate normal distribution, in such a way that the resulting structure could be considered to consist of natural clusters that exhibit the properties of external isolation and internal cohesion. The details about the centers and widths of these normal distributions are precisely as described in Ref. [27].

Following Milligan and Cooper's procedure, the design factors corresponding to the number of clusters and to the number of dimensions were crossed with each other and both were crossed with a third factor that determines the number of objects within the clusters. Provided that the number of objects in each data set is fixed, this third factor directly affects not only the cluster densities, but the overall data balance as well. This factor consists of three levels, where one level corresponds to an equal number of objects in each cluster (or as close to equality as possible), the second level requires that one cluster must always contain 10% of the data objects, whereas the third level requires that one cluster must contain 60% of the objects. The remaining objects were distributed as evenly as possible across the other clusters present in the data. Overall, there were 36 cells in the design (4 numbers of clusters × 3 dimensions × 3 balances). Three sampling replications were generated for each cell, thus producing a total of 108 data sets.

5.2. Experimental Methodology

With the data sets in hand, four versions of the standard agglomerative hierarchical clustering algorithm [1], namely, *single linkage*, *complete linkage*, *average linkage*, and *Ward's*, were systematically applied to each data set. For each data set, each algorithm produced a hierarchy of data partitions with the number of clusters ranging from $k = 2$ through $k = N = 50$. Such partitions can then be evaluated by the relative clustering validity criteria under investigation. For illustration purposes, the evaluation results produced by four optimization-like criteria and two difference-like criteria (with the corresponding optimization-like transformations) for partitions of a typical data set with $k^* = 5$ are displayed in Figs. 6 and 7, respectively.

All the criteria in Fig. 6 as well as those transformed ones in Fig. 7 exhibit a primary (maximum or minimum) peak at the expected number of clusters, $k^* = 5$. However, it can be observed in Fig. 6 that some criteria exhibit a secondary peak when evaluating partitions with too many clusters, particularly as k approaches the number of objects,

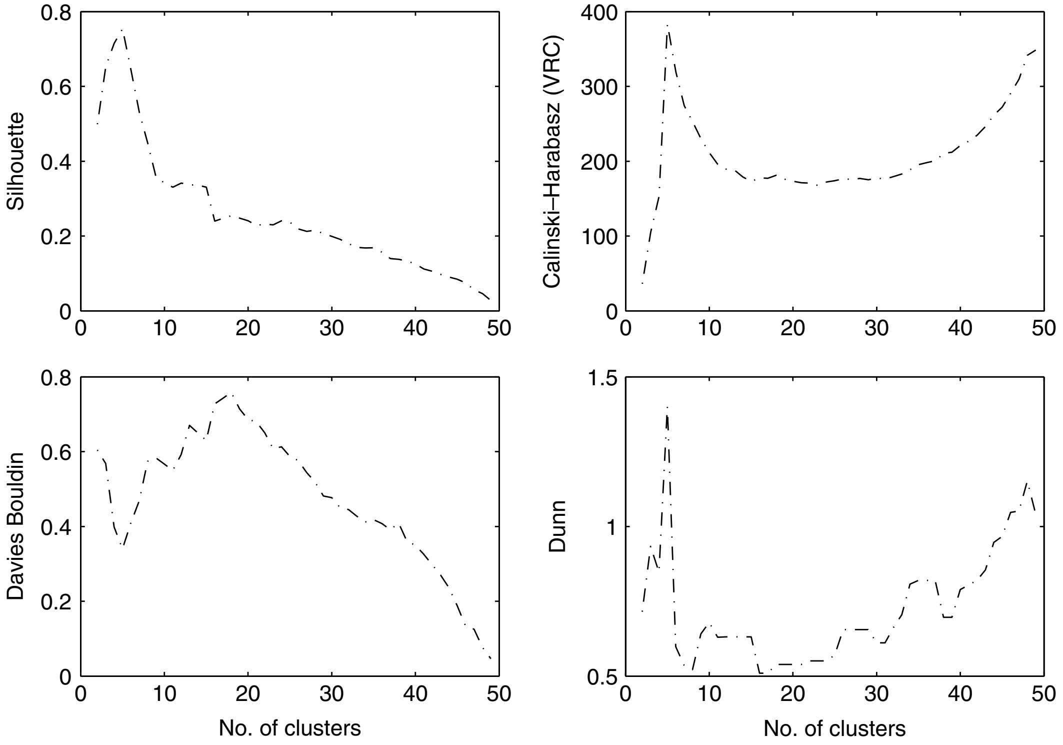


Fig. 6 Values of four optimization-like relative clustering validity criteria: hierarchical partitions of a typical data set for $k = 2, \ldots, 50$.

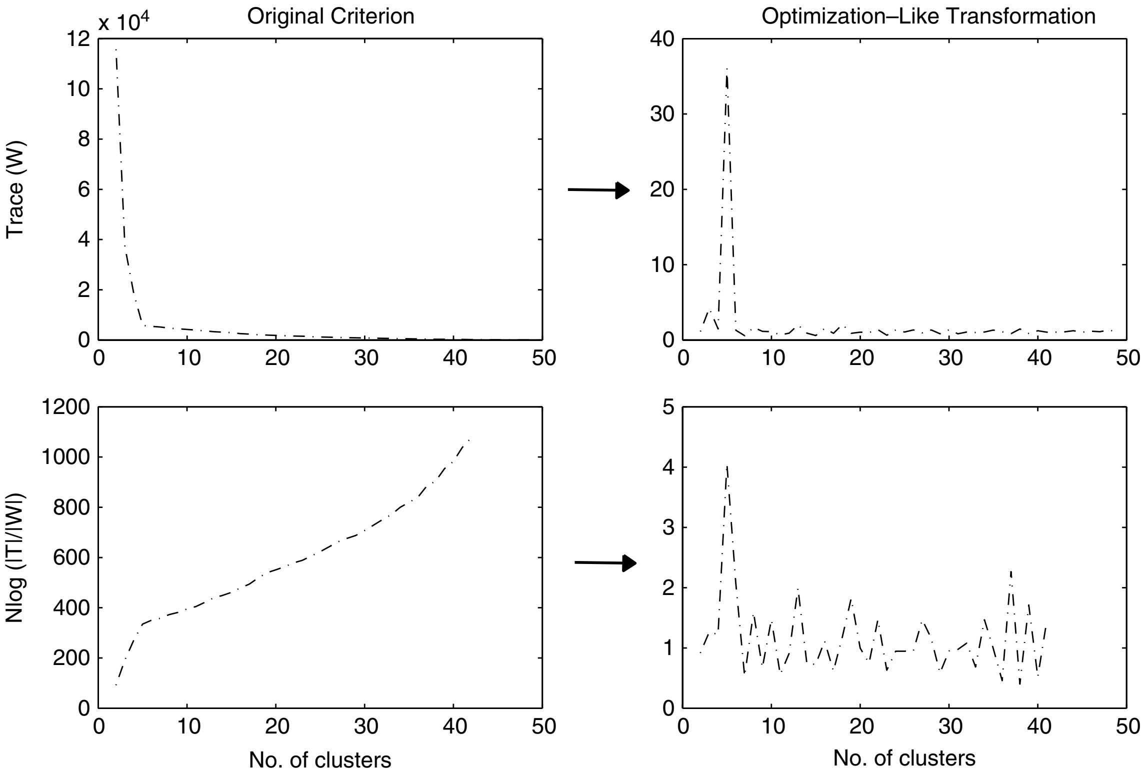


Fig. 7 Values of two difference-like criteria (left) and their corresponding optimization-like transformations (right): hierarchical partitions of a typical data set for $k = 2, \ldots, 50$.

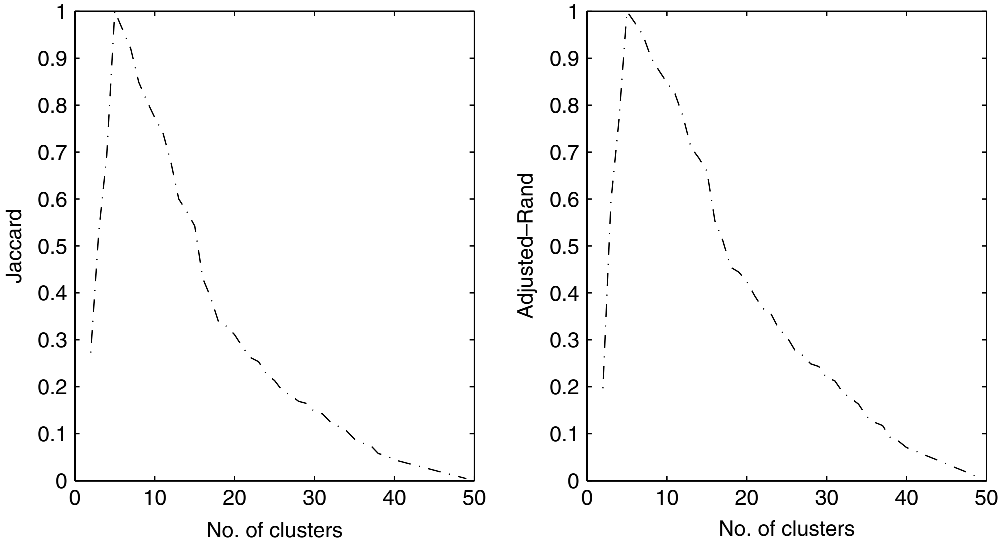


Fig. 8 Values of the Jaccard coefficient and ARI: hierarchical partitions of a typical data set for $k = 2, \ldots, 50$.

$N = 50$. Such a behavior is also observed in other relative clustering validity criteria[5] and must be carefully dealt with when assessing their performances. The reason is that this eventual secondary peak, which may be less or more intense depending on the criterion and on the data set as well, can clearly affect the correlation between the relative and external validity values. In addition, the values of some criteria may not be defined for partitions with too many clusters. For instance, the criterion Nlog(|T|/|W|) is no longer defined for values of k beyond 42 in Fig. 7 due to the numerical singularity of matrix $\mathbf{W}$ when k approaches N. Fortunately, this is not a serious situation inasmuch as partitions with k approaching N are completely out of scope in practical clustering analysis. In order to prevent problems, it is recommended to perform the methodology proposed in Section 4 over a set of partitions with the number of clusters limited within an acceptable range. Such an acceptable range depends on the application domain, but there are some general rules that can be useful to guide the less experienced user in practical situations. One rule of thumb is to set the upper bound of this interval to $k_{\max} = \sqrt{N}$ or so (≈ 8 for $N = 50$). Another possibility, which can be particularly useful if the number of objects is small, is to set this upper bound to $k_{\max} = N/2$. The rationale behind such a more conservative number is that, conceptually speaking, a cluster is expected to be composed of at least two objects, otherwise being an outlier rather than a group of objects. Both these intervals, i.e. $k \in \{2, \ldots, 8\}$ and $k \in \{2, \ldots, 25\}$, will be adopted in this part of the study. This means that only a subset of the partitions produced by the hierarchical algorithms for each data set will be used in the experiments, namely, 3024 partitions for the first interval (4 algorithms × 108 data sets × 7 values for k) and 10, 368 partitions for the second interval ($4 \times 108 \times 24$). Having such selected partitions in hand, the relative and external validity criteria can be computed for all of them and the comparison procedure described in Section 4 can be accomplished.

Remark 3. In order to compute difference-like criteria for partitions with $k = 2$ and $k = k_{\max}$ using the optimization-like transformation in Eq. (37), it is actually necessary to compute the original indexes for $k = 1$ and $k = k_{\max} + 1$. Two of the difference-like criteria reviewed in Section 2.2, however, are not defined for $k = 1$, namely, log(SSB/SSW) and McClain–Rao. As a consequence, the transformed versions of these criteria cannot be assessed for partitions with $k = 2$.

For better confidence of the analysis, the results corresponding to two distinct external validity criteria described in Section 3 are reported here; namely, the ARI and the Jaccard coefficient. The values of these criteria for the set of partitions produced by a hierarchical algorithm applied to a typical data set with $k^* = 5$ are illustrated in Fig. 8. This figure shows that the behaviors of ARI and Jaccard are in conformity with each other and, accordingly, these external indexes can indeed be used together for better confidence of the analysis. Similarly, two distinct correlation indexes are also used to measure the agreement level between the relative and external criteria (please, refer to step 3 of the comparison procedure in Section 4). One of the indexes is the classic Pearson coefficient. Although Pearson is recognizably powerful, it is also well-known to be somewhat

[5] The value of PBM described in Section 2.1.9, for example, goes to infinity as $k \to N$, because E_k in Eq. (21) tends to zero.

sensitive to dominant peaks in the sequences of values under evaluation. In order to prevent the results from being biased by the subset of best relative and external validity values, probably associated with partitions around k^*, an additional correlation index is also adopted here that can be more sensitive to those nonpeak values of the sequences. It is a weighted version of the Goodman-Kruskal index [7,53], named Weighted-Goodman-Kruskal (WGK), which may be particularly appropriate to clustering analysis for it is fully sensitive to both the ranks and the magnitude proportions of the sequences under evaluation [35].

Histograms of the results strongly suggest that the observed sample distributions hardly satisfy the normality assumption. For this reason, outcomes of parametric statistical tests will not be reported here. Instead, the well-known Wilcoxon/Mann-Whitney (W/M-W) test will be adopted. The efficiency of the W/M-W test is 0.95 with respect to parametric tests like the t-test or the z-test even if the data are normal. Thus, even when the normality assumption is satisfied, the W/M-W test might be preferred [54]. This test will be used in this work to compare the results of every pair of relative validity criteria as two sampled populations. In addition, another test will also be applied that subdivides these sampled populations into blocks. In this case, each block is treated as an independent subsample composed of those instances that are related to data sets generated from a particular configuration of the data generator. A particular configuration of the data generator corresponds precisely to one of those 36 design cells composed of data sets with the same numbers of clusters, dimensions, and the same balance. This is called *two-way randomized block design*, one way referring to samples coming from different relative criteria and another way referring to samples coming from different configurations of the data generator. A nonparametric test of this class is named Friedman test[6], which will be adopted here to reinforce the results of the W/M-W test in a conservative (duplicated) manner.

Finally, for the sake of comparison, the traditional methodology adopted by Milligan and Cooper [14] will also be carried out here for all those 40 relative validity criteria reviewed in Section 2. Let us recall that, in the traditional methodology, each criterion evaluates the set of partitions produced by a given hierarchical algorithm (*single linkage*, *complete linkage*, *average linkage*, and *Ward's*, for $k = 2, \ldots, k_{\max}$)[7] and, then, the number of clusters in the partition elected as the best one by the criterion is compared against the *rigth* number of clusters known to exist in that particular data set. The number of hits achieved over the whole collection of data sets and different hierarchical algorithms is taken as the final performance of that criterion. The maximum number of hits is 432 (108 data sets × 4 hierarchical algorithms).

5.3. Results and Discussions

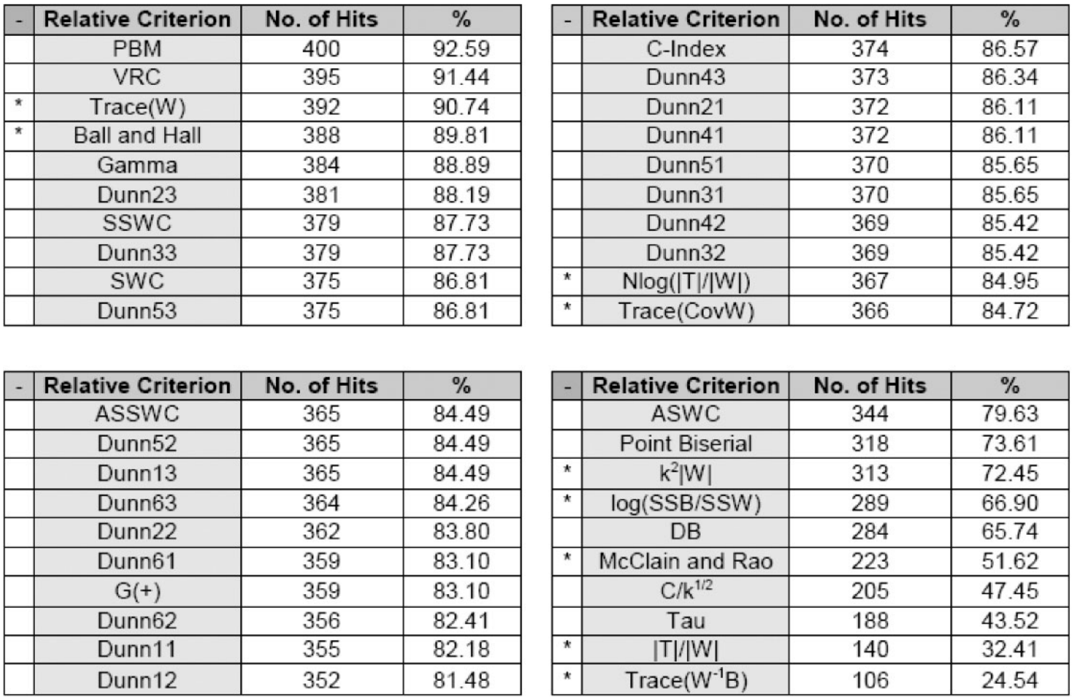
The final results obtained by the traditional methodology are illustrated in Fig. 9. Before proceeding with the discussions, note that, following the terminology adopted in Ref. [21], the variants of Dunn's index have been named Dunn*XY*. More specifically, $X = 1$ refers to the original definition of *set distance* (see Section 2.1.3), whereas $X \in \{2, \ldots, 6\}$ refers to the alternative definitions given by Eqs. (11), (12), (13), (14), and (15), respectively. Similarly, $Y = 1$ refers to the original definition of *diameter*, whereas $Y \in \{2, 3\}$ refers to Eqs. (16) and (17), respectively. Note that, following this terminology, the original index is named Dunn11.

The results achieved by the optimization-like criteria that had already been considered in the original study by Milligan and Cooper [14] were, as expected, similar to those reported in that reference. Accordingly, the corresponding findings and conclusions still remain valid. Considering those optimization-like criteria not covered by Milligan and Cooper [14], it is worth highlighting the excellent performance provided by PBM, which correctly elected the best partition in 92.59% of the assessed scenarios (400 out of 432). It is also worth mentioning that the variants of Dunn provided good relative performances when compared to the original criterion (whose number of hits is 355, i.e. 82.18%). On the other hand, the alternative variants of the silhouette (ASWC and ASSWC) showed slightly worse performances when compared to both the original criterion (SWC) and its simplified version (SSWC). This suggests that such alternative variants may have a slightly inferior capacity of indicating as the best partition one with the expected number of clusters (at least with respect to data sets similar to those considered in the present study).

By comparing the results shown in Fig. 9 with those reported in Ref. [14], one can observe that all difference-like criteria have performed better when Eq. (37) (proposed here) is used to transform them into optimization-like criteria. In this context, we would like to highlight the superior performances of trace(W) and Ball–Hall, which got 90.74 and 89.81% of hits, respectively (in Ref. [14], the corresponding results are 27.78 and 29.63%, respectively). These criteria are among the simplest ones (both conceptually and computationally), because they only need information on the distances between objects and centroids of their clusters. The criteria Nlog(|T|/|W|) and trace(CovW) have shown competitive performances as well, namely, 84.95 and 84.72% of hits, respectively (against 34.49 and 28.01% reported in Ref. [14]). The remaining difference-like criteria

[6] A relative of the two-way analysis of variance (ANOVA) test.

[7] $k_{\max}$ will be set to 25, which is precisely the same upper bound adopted for the alternative comparison methodology.



-	Relative Criterion	No. of Hits	%
	PBM	400	92.59
	VRC	395	91.44
*	Trace(W)	392	90.74
*	Ball and Hall	388	89.81
	Gamma	384	88.89
	Dunn23	381	88.19
	SSWC	379	87.73
	Dunn33	379	87.73
	SWC	375	86.81
	Dunn53	375	86.81

-	Relative Criterion	No. of Hits	%
	C-Index	374	86.57
	Dunn43	373	86.34
	Dunn21	372	86.11
	Dunn41	372	86.11
	Dunn51	370	85.65
	Dunn31	370	85.65
	Dunn42	369	85.42
	Dunn32	369	85.42
*	Nlog(|T|/|W|)	367	84.95
*	Trace(CovW)	366	84.72

-	Relative Criterion	No. of Hits	%
	ASSWC	365	84.49
	Dunn52	365	84.49
	Dunn13	365	84.49
	Dunn63	364	84.26
	Dunn22	362	83.80
	Dunn61	359	83.10
	G(+)	359	83.10
	Dunn62	356	82.41
	Dunn11	355	82.18
	Dunn12	352	81.48

-	Relative Criterion	No. of Hits	%
	ASWC	344	79.63
	Point Biserial	318	73.61
*	k^2|W|	313	72.45
*	log(SSB/SSW)	289	66.90
	DB	284	65.74
*	McClain and Rao	223	51.62
	C/$k^{1/2}$	205	47.45
	Tau	188	43.52
*	|T|/|W|	140	32.41
*	Trace($W^{-1}B$)	106	24.54

Fig. 9 Number of hits for each criterion considering a collection of 108 data sets and partitions found by four hierarchical algorithms (432 combinations/scenarios)—Difference-like criteria are marked with an asterisk (*).

have also shown better results when compared with those reported by Milligan and Cooper, but not when compared with other relative validity criteria included into the present analysis. In particular, note that |T|/|W| and trace($W^{-1}B$) have achieved only 32.41 and 24.54% of hits, respectively (against 0.00 and 19.44% in Ref. [14]).

Using the alternative methodology discussed in Section 4 (when the Pearson coefficient, the Jaccard criterion, and $k_{\max} = 25$ are employed), we found the results reported in Fig. 10. Before proceeding with the discussions, note that, for the sake of compactness and clarity, Fig. 10 displays only 3 out of the 17 variants of the original Dunn's index (Dunn11) that were reviewed in Section 2.1.4, namely, Dunn12, Dunn13, and Dunn62. These variants are those that have shown the best overall performances in this first part of the experiments.

The mean of the correlation values between each relative criterion and the external criterion (Jaccard)—computed over the whole collection of $N_D = 108$ available values (one per data set)—is displayed at the bottom bar of Fig. 10. In this figure, the relative validity criteria have been placed in the rows and columns of the top table according to the ordering established by their correlation means (decreasing order from top to bottom and from left to right). The value displayed in each cell of the top table corresponds to the difference between the correlation means of the corresponding pair of relative criteria. A shaded cell indicates that the corresponding difference is statistically significant at the $\alpha = 5\%$ level (one-sided test). Darker shaded cells indicate that significance has been observed with respect to both W/M-W and Friedman tests, whereas lighter shaded cells denote significance with respect to one of these tests only.

The analyses exemplified in Fig. 10 have also been carried out for every possible scenario under evaluation, that is to say, for all combinations of: two correlation measures (WGK and Pearson), two external criteria (Jaccard and ARI), and two upper bounds for the number of clusters ($k_{\max} = 8$ and $k_{\max} = 25$), thus resulting in eight different tables similar to the one depicted in Fig. 10. From such a complete set of tables[8] (omitted here for the sake of compactness) one can derive some interesting conclusions, summarized below:

(1) In general, optimization-like criteria exhibited results that are better than those provided by difference-like criteria. In several evaluation scenarios, this is valid with strong statistical significance.

(2) The optimization-like criteria that exhibited the best performances were point-biserial, tau, ASWC, ASSWC, PBM, SWC, SSWC, and VRC. These criteria provided results that are better than those obtained

[8] Available at http://www.icmc.usp.br/∼campello/Sub_Pages/Selected_Publications.htm.

		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Point-Biserial	**A**	0.000	0.046	0.226	0.247	0.262	0.289	0.306	0.373	0.390	0.408	0.488	0.555	0.566	0.571	0.584	0.636	0.642	0.645	0.694	0.705	0.729	0.736	0.768	0.822	0.837	1.107
Tau	**B**	-0.046	0.000	0.180	0.201	0.216	0.243	0.260	0.327	0.344	0.362	0.442	0.509	0.520	0.525	0.538	0.590	0.597	0.599	0.649	0.659	0.683	0.690	0.722	0.776	0.791	1.061
$C/k^{1/2}$	**C**	-0.226	-0.180	0.000	0.021	0.036	0.063	0.080	0.147	0.164	0.182	0.263	0.329	0.340	0.345	0.358	0.410	0.417	0.419	0.469	0.479	0.504	0.510	0.542	0.596	0.611	0.881
ASWC	**D**	-0.247	-0.201	-0.021	0.000	0.015	0.042	0.060	0.127	0.143	0.161	0.242	0.308	0.319	0.324	0.338	0.390	0.396	0.398	0.448	0.458	0.483	0.489	0.521	0.575	0.590	0.860
ASSWC	**E**	-0.262	-0.216	-0.036	-0.015	0.000	0.027	0.045	0.112	0.128	0.146	0.227	0.293	0.304	0.309	0.323	0.375	0.381	0.384	0.433	0.443	0.468	0.474	0.506	0.560	0.575	0.846
PBM	**F**	-0.289	-0.243	-0.063	-0.042	-0.027	0.000	0.017	0.084	0.101	0.119	0.199	0.266	0.277	0.282	0.295	0.347	0.353	0.356	0.406	0.416	0.440	0.447	0.479	0.533	0.548	0.818
SWC	**G**	-0.306	-0.260	-0.080	-0.060	-0.045	-0.017	0.000	0.067	0.083	0.102	0.182	0.249	0.260	0.265	0.278	0.330	0.336	0.339	0.388	0.399	0.423	0.430	0.462	0.516	0.530	0.801
SSWC	**H**	-0.373	-0.327	-0.147	-0.127	-0.112	-0.084	-0.067	0.000	0.016	0.035	0.115	0.181	0.193	0.198	0.211	0.263	0.269	0.272	0.321	0.332	0.356	0.363	0.395	0.449	0.463	0.734
Dunn12	**I**	-0.390	-0.344	-0.164	-0.143	-0.128	-0.101	-0.083	-0.016	0.000	0.018	0.099	0.165	0.176	0.181	0.195	0.247	0.253	0.255	0.305	0.315	0.340	0.346	0.378	0.432	0.447	0.717
Dunn62	**J**	-0.408	-0.362	-0.182	-0.161	-0.146	-0.119	-0.102	-0.035	-0.018	0.000	0.080	0.147	0.158	0.163	0.176	0.228	0.234	0.237	0.287	0.297	0.321	0.328	0.360	0.414	0.429	0.699
Dunn13	**K**	-0.488	-0.442	-0.263	-0.242	-0.227	-0.199	-0.182	-0.115	-0.099	-0.080	0.000	0.066	0.078	0.082	0.096	0.148	0.154	0.157	0.206	0.217	0.241	0.248	0.280	0.334	0.348	0.619
VRC	**L**	-0.555	-0.509	-0.329	-0.308	-0.293	-0.266	-0.249	-0.181	-0.165	-0.147	-0.066	0.000	0.011	0.016	0.030	0.082	0.088	0.090	0.140	0.150	0.175	0.181	0.213	0.267	0.282	0.552
Ball and Hall	**M**	-0.566	-0.520	-0.340	-0.319	-0.304	-0.277	-0.260	-0.193	-0.176	-0.158	-0.078	-0.011	0.000	0.005	0.018	0.070	0.076	0.079	0.129	0.139	0.163	0.170	0.202	0.256	0.271	0.541
Trace(W)	**N**	-0.571	-0.525	-0.345	-0.324	-0.309	-0.282	-0.265	-0.198	-0.181	-0.163	-0.082	-0.016	-0.005	0.000	0.013	0.065	0.072	0.074	0.124	0.134	0.159	0.165	0.197	0.251	0.266	0.536
DB	**O**	-0.584	-0.538	-0.358	-0.338	-0.323	-0.295	-0.278	-0.211	-0.195	-0.176	-0.096	-0.030	-0.018	-0.013	0.000	0.052	0.058	0.061	0.110	0.121	0.145	0.152	0.184	0.238	0.252	0.523
Nlog(|T|/|W|)	**P**	-0.636	-0.590	-0.410	-0.390	-0.375	-0.347	-0.330	-0.263	-0.247	-0.228	-0.148	-0.082	-0.070	-0.065	-0.052	0.000	0.006	0.009	0.058	0.069	0.093	0.100	0.132	0.186	0.200	0.471
Trace(CovW)	**Q**	-0.642	-0.597	-0.417	-0.396	-0.381	-0.353	-0.336	-0.269	-0.253	-0.234	-0.154	-0.088	-0.076	-0.072	-0.058	-0.006	0.000	0.003	0.052	0.063	0.087	0.094	0.126	0.180	0.194	0.465
k^2|W|	**R**	-0.645	-0.599	-0.419	-0.398	-0.384	-0.356	-0.339	-0.272	-0.255	-0.237	-0.157	-0.090	-0.079	-0.074	-0.061	-0.009	-0.003	0.000	0.049	0.060	0.084	0.091	0.123	0.177	0.192	0.462
log(SSB/SSW)	**S**	-0.694	-0.649	-0.469	-0.448	-0.433	-0.406	-0.388	-0.321	-0.305	-0.287	-0.206	-0.140	-0.129	-0.124	-0.110	-0.058	-0.052	-0.049	0.000	0.010	0.035	0.041	0.074	0.128	0.142	0.413
Dunn11	**T**	-0.705	-0.659	-0.479	-0.458	-0.443	-0.416	-0.399	-0.332	-0.315	-0.297	-0.217	-0.150	-0.139	-0.134	-0.121	-0.069	-0.063	-0.060	-0.010	0.000	0.024	0.031	0.063	0.117	0.132	0.402
Gamma	**U**	-0.729	-0.683	-0.504	-0.483	-0.468	-0.440	-0.423	-0.356	-0.340	-0.321	-0.241	-0.175	-0.163	-0.159	-0.145	-0.093	-0.087	-0.084	-0.035	-0.024	0.000	0.007	0.039	0.093	0.107	0.378
McClain and Rao	**V**	-0.736	-0.690	-0.510	-0.489	-0.474	-0.447	-0.430	-0.363	-0.346	-0.328	-0.248	-0.181	-0.170	-0.165	-0.152	-0.100	-0.094	-0.091	-0.041	-0.031	-0.007	0.000	0.032	0.086	0.101	0.371
C-Index	**W**	-0.768	-0.722	-0.542	-0.521	-0.506	-0.479	-0.462	-0.395	-0.378	-0.360	-0.280	-0.213	-0.202	-0.197	-0.184	-0.132	-0.126	-0.123	-0.074	-0.063	-0.039	-0.032	0.000	0.054	0.069	0.339
|T|/|W|	**X**	-0.822	-0.776	-0.596	-0.575	-0.560	-0.533	-0.516	-0.449	-0.432	-0.414	-0.334	-0.267	-0.256	-0.251	-0.238	-0.186	-0.180	-0.177	-0.128	-0.117	-0.093	-0.086	-0.054	0.000	0.015	0.285
Trace($W^{-1}B$)	**Y**	-0.837	-0.791	-0.611	-0.590	-0.575	-0.548	-0.530	-0.463	-0.447	-0.429	-0.348	-0.282	-0.271	-0.266	-0.252	-0.200	-0.194	-0.192	-0.142	-0.132	-0.107	-0.101	-0.069	-0.015	0.000	0.270
G(+)	**Z**	-1.107	-1.061	-0.881	-0.860	-0.846	-0.818	-0.801	-0.734	-0.717	-0.699	-0.619	-0.552	-0.541	-0.536	-0.523	-0.471	-0.465	-0.462	-0.413	-0.402	-0.378	-0.371	-0.339	-0.285	-0.270	0.000
	Mean	0.959	0.913	0.733	0.712	0.697	0.670	0.653	0.586	0.569	0.551	0.471	0.404	0.393	0.388	0.375	0.323	0.316	0.314	0.264	0.254	0.230	0.223	0.191	0.137	0.122	-0.148

Fig. 10 Mean values (bottom bar) and their differences (cells) for Pearson correlation between relative and external (Jaccard) criteria: $k_{\max} = 25$.

by *all* difference-like criteria and by the original Dunn's index as well, with statistical significance.

(3) The different versions of the silhouette criterion presented comparable results. Then, once the simplified versions (SSWC and ASSWC) have lower computational requirements, they may be preferred, especially when dealing with large data sets.

(4) In general, Dunn12 and Dunn13 outperformed the original criterion (Dunn11), particularly in those scenarios for which $k_{\max} = 25$.

(5) The point-biserial criterion provided superior results (with statistical significance) when compared with *all* criteria, except gamma in the specific scenario involving Pearson correlation and $k_{\max} = 8$.

It is important to remark that differences were observed between the performances of some criteria when switching from $k_{\max} = 8$ to $k_{\max} = 25$. In particular, expressive drops in performance were noticed when $k_{\max} = 25$. This suggests that the corresponding criteria are not robust to keep working accurately in the presence of bad quality partitions—in this case formed by numbers of clusters quite different from[9] k^*. As such, these criteria may not be recommended for real-world applications involving complicating factors such as noisy data, overlapping clusters, high dimensionality, among others. This is the case of G(+), gamma, and C-index, whose performance losses took place with respect to all correlation measures (Pearson/WGK) and external criteria (Jaccard/ARI) when $k_{\max} = 25$ (for illustration purposes, scatter plots of G(+) versus Jaccard for a typical data set are depicted in Figs. 11 and 12 for $k_{\max} = 8$ and $k_{\max} = 25$, respectively). Curiously, $C/\sqrt{k}$ behaved in the opposite way, giving better results when $k_{\max} = 25$. A detailed and careful analysis of such a behavior does not favor the criterion. As observed in Section 2.1.15, $C/\sqrt{k}$ may be dominated by the number of clusters, when this quantity is large enough. More precisely, $C/\sqrt{k} \rightarrow 1/\sqrt{k}$ for large values of k. Such a decreasingly monotonic behavior of this criterion as a function of k is analogous to the one exhibited by some external criteria when the data set in hand has a few clusters (small k^*), which is precisely the case here addressed. This explains the misleading performance of $C/\sqrt{k}$ for $k_{\max} = 25$ and reinforces the need for a careful

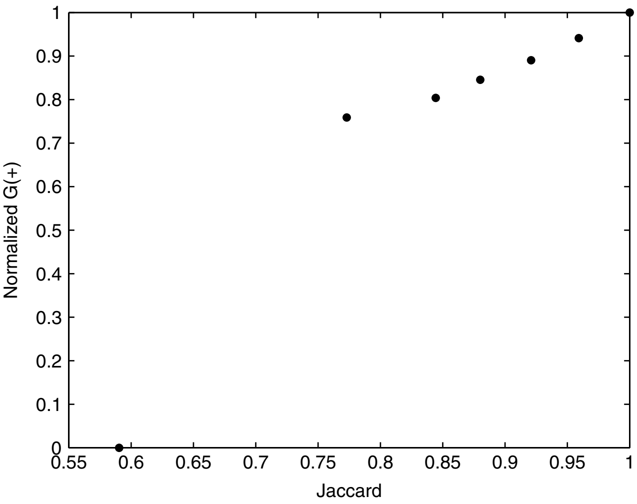


Fig. 11 Scatter plot of normalized values of G(+) versus Jaccard for partitions of a typical data set: $k_{\max} = 8$.

[9] It is worth noticing that $k^* = 2$, 3, 4, or 5 in the experiments reported in this section.

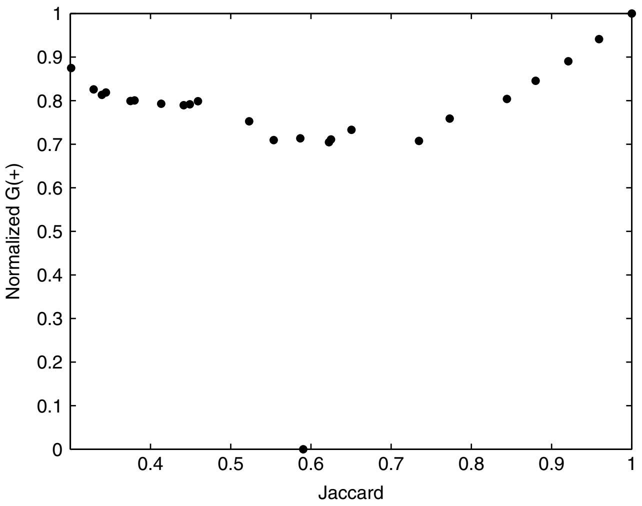


Fig. 12 Scatter plot of normalized values of G(+) versus Jaccard for partitions of a typical data set: $k_{\max} = 25$.

design of the experiments and an attentive evaluation of the results. Discussions on experiments especially designed to detect biases with respect to the number of clusters, such as the one just described, are covered in the second part of the empirical evaluation (Section 6).

Comparative analyses of the results obtained using the different (traditional and alternative) methodologies adopted in this first part of the experiments were carried out with the aim of detecting persisting and conflicting patterns in the performance of the criteria under investigation. From this standpoint, the main conclusions that can be derived are:

(1) In absolute terms, trace(W), trace(CovW), Nlog (|T|/|W|), Ball–Hall, and Dunn11 provided reasonable results when the traditional methodology is employed. However, their performances are poor from the perspective of the alternative methodology. This suggests that these criteria can estimate the number of clusters satisfactorily, but they are not accurate when evaluating the quality of a collection of partitions in relative terms (with respect to the information provided by external criteria). Opposite behavior was observed from point-biserial and tau.

(2) The criteria for which performances were superior for both methodologies (being among the top ten in both cases) are: SWC, SSWC, VRC, and PBM.

It is worth stressing that the results discussed in this section were obtained using data partitions found by hierarchical algorithms applied to data sets formed by a small number of objects ($N = 50$), a few clusters ($k^* \in \{2, 3, 4, 5\}$), and a small number of attributes ($n \in \{4, 6, 8\}$). Aimed at investigating the behavior of the relative validity criteria in a broader context, the following section presents a complementary experimental study.

6. EXPERIMENTAL RESULTS—PART II

In this second part of our empirical evaluation, we report the results of more comprehensive experiments, using larger data sets formed by more diversified numbers of clusters and attributes.

6.1. Data Sets

Milligan and Cooper used data sets formed by 50 objects, possibly due to past computational limitations. As previously addressed, they used data sets formed by three different numbers of dimensions ($n \in \{4, 6, 8\}$) and four different numbers of clusters ($k^* \in \{2, 3, 4, 5\}$). These were also used here in this paper, in the first part of the experiments (Section 5). In this second part of the experiments, data sets formed by a number of objects ten times more representative ($N = 500$) have been adopted. The number of dimensions and clusters have also been expanded in order to obtain higher diversity in relation to these characteristics, as well as to allow the study on how these factors can affect the performance of relative validity criteria. Considering the number of dimensions, we generated two subcategories of data sets: one of them formed by fewer dimensions ($n \in \{2, 3, 4\}$) and the other one formed by more dimensions ($n \in \{22, 23, 24\}$). Analogously, two subcategories were conceived for the number of clusters ($k^* \in \{2, 4, 6\}$ and $k^* \in \{12, 14, 16\}$). To do so, the same data set generator described in Section 5.1 was employed.

Following the procedure adopted in Section 5.1, the experimental design factors corresponding to the number of clusters and dimensions were combined and crossed with a third factor that determines the quantity of objects in each cluster. This third factor consists of three levels. The first level originates approximately balanced clusters. In the second level, one of the clusters has 10% of the objects, whereas the remaining objects are approximately balanced among the other clusters. In the third level, one of the clusters has 60% of the objects, and the remaining objects are again approximately balanced among the other clusters. However, this level may cause difficulties for data sets formed by many clusters, for that it implies in forming many clusters with just a few objects, thus making it harder for a clustering algorithm to find a subset of good quality partitions. For this reason, the third level has been changed for data sets formed by $k^* \in \{12, 14, 16\}$ clusters, in such a way that one of the clusters contains 20% (instead of 60%)

of the objects. In brief, the three experimental design factors described above have been combined, thus producing a set of 108 (6 number of clusters × 6 dimensions × 3 balances) experimental design cells to generate data sets. For better statistical confidence, nine independent replications were used for each cell, thus resulting in a collection of 972 data sets.

Remark 4. Noisy data sets were not considered in the experiments because the presence of noise could introduce a confounding effect into the analyses of the behavior of the validity criteria under investigation, namely, the behavior of the clustering algorithms used to generate the data partitions when they face noisy or noiseless data. However, notice that, even in the absence of noise, the performances of the validity criteria were evaluated when assessing partitions of varied qualities. Indeed, a large amount of partitions for each data set was generated that includes not only good (possibly optimal) partitions, but bad partitions as well, namely: (i) partitions with the number of clusters very different from the number of natural clusters known to exist in each the data set (k^*); and (ii) partitions that may represent local optima of the clustering algorithm adopted in this part of the experiments (see the next section).

6.2. Experimental Methodology

Following Milligan and Cooper's methodology [14,27], in the first part of this study (Section 5) we have adopted hierarchical clustering algorithms to perform all the experiments. In this part of the experiments, the classic k-means algorithm [10] is adopted in lieu of hierarchical algorithms, for the following main reasons: (i) k-means has been elected and listed among the top ten most influential data mining algorithms [55], possibly because it is both very simple and quite scalable. Indeed, in contrast to the squared asymptotic running time of hierarchical algorithms with respect to the number of objects (N), k-means has linear time complexity with respect to any aspect of the problem size [55]; (ii) The hierarchy produced by a hierarchical algorithm provides only a single partition of the data set for every value of k. For better confidence of the results, it is preferred here to produce a number of partitions for every k by running the k-means algorithm from different initializations of prototypes (as randomly selected data objects). The use of a number of partitions for every k allows increasing diversity of the collection of partitions. The evaluation of such a more diverse collection of partitions results in a more reliable set of N_π clustering validity values associated with a given data set—refer to steps 2 and 3 of the comparison procedure described in Section 4. In particular, 20 partitions are obtained for every k ranging from $k = 2$ through $k = k_{\max}$, which represents an amount of $(k_{\max} - 1) \times 20$ partitions for each data set.

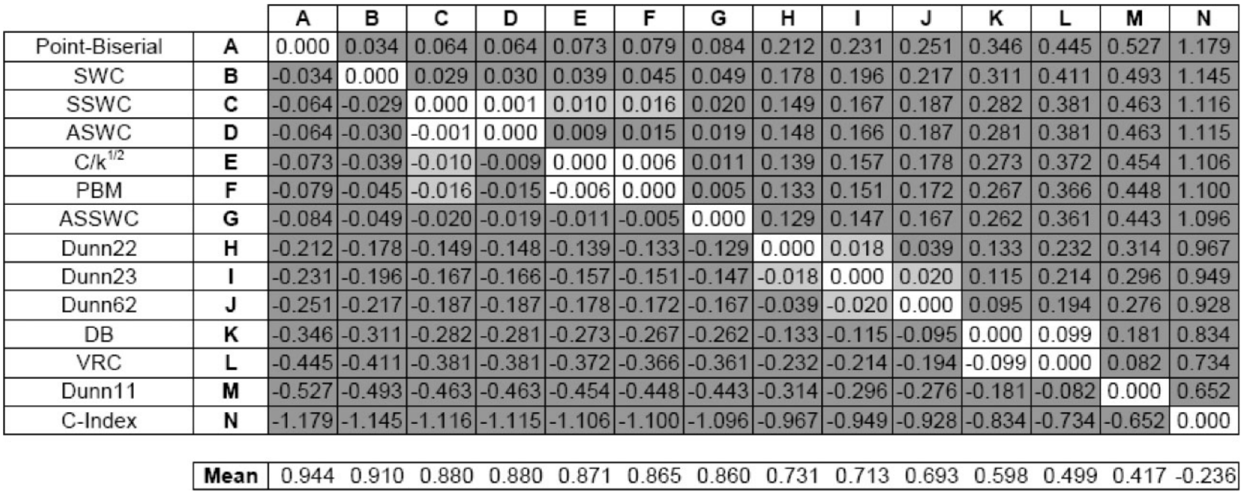
As in Section 5, two distinct evaluation scenarios with respect to different values of $k_{\max}$ are considered here. One of them is $k_{\max} = \sqrt{N}$ (≈ 23 for $N = 500$). As this value is close to the number of clusters in some of the data sets adopted in this part of the experiments, k^*, we also here assess partitions formed by $k_{\max} = 50$. This value allows evaluating partitions with a number of clusters, k, significantly different from k^*.

In brief, k-means is repeatedly applied to each data set with k varying from 2 to $k_{\max}$, where $k_{\max} = 23$ or $k_{\max} = 50$. For each k, k-means runs 20 times, with different initial prototypes. Thus, the algorithm produces $N_\pi = (50 - 1) \times 20 = 980$ partitions for each data set, among which only $N_\pi = (23 - 1) \times 20 = 440$ partitions are actually used in the evaluation scenario with $k_{\max} = 23$. Since there are 972 available data sets, a collection of $972 \times 980 = 952,560$ partitions has been obtained. For each partition, the relative and external validity criteria can be computed and the procedure described in Section 4 can be performed. To do so, the 972 data sets are subdivided into four subcategories, each of which formed by 243 data sets and characterized by fewer or more dimensions (either $n \in \{2, 3, 4\}$ or $n \in \{22, 23, 24\}$) and clusters (either $k^* \in \{2, 4, 6\}$ or $k^* \in \{12, 14, 16\}$). The comparison procedure described in Section 4 is then performed individually for each of these subcategories in order to allow evaluating the influence of the factors n and k^* on the behavior of the relative validity criteria.

Owing to the very high computational complexity of the relative criteria known as gamma (Section 2.1.11), G(+) (Section 2.1.12), and tau (Section 2.1.13), they could hardly be used in real-world applications, especially those involving data sets with "large" N. For this reason, they have not been assessed in the current experiments. Moreover, difference-like criteria (Section 2.2) have not been included into the current analyses as well, for they require a (hierarchical) relation of succession between partitions with consecutive values of k. This does not happen from multiple, independent runs of k-means.

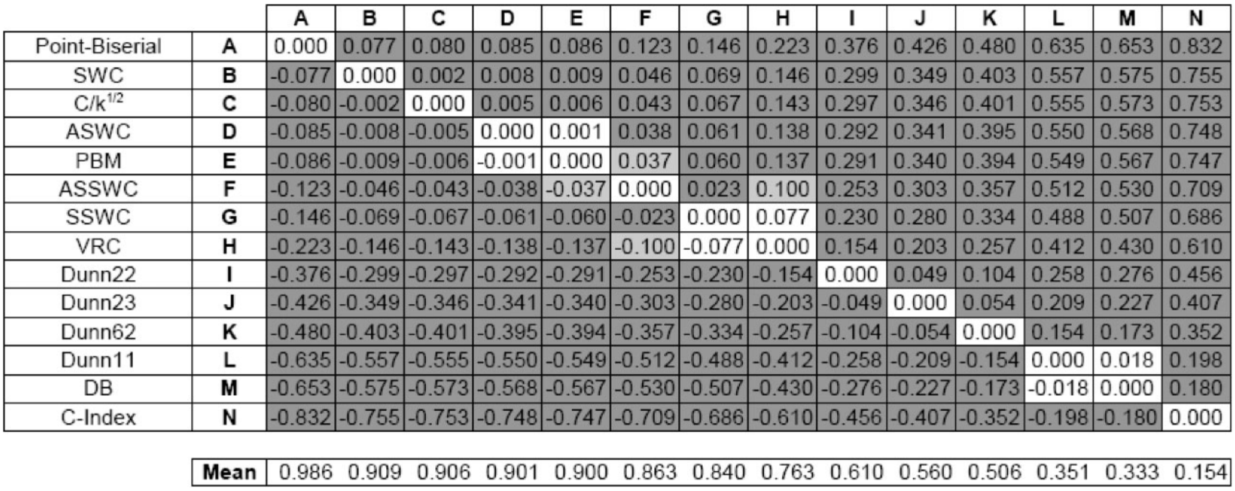
6.3. Results and Discussions

The final results obtained by means of the methodology described in Section 4 using the Pearson correlation coefficient, the Jaccard external criterion, and $k_{\max} = 50$, are presented in Figs. 13–17. The presentation scheme and layout of these figures is precisely the same as that adopted in Fig. 10. Figures 13–16 correspond to the results achieved separately from each of the four subcategories (formed by 243 data sets) involving combinations of more (or fewer) dimensions and clusters. Figure 17 reports the



		A	B	C	D	E	F	G	H	I	J	K	L	M	N
Point-Biserial	A	0.000	0.034	0.064	0.064	0.073	0.079	0.084	0.212	0.231	0.251	0.346	0.445	0.527	1.179
SWC	B	-0.034	0.000	0.029	0.030	0.039	0.045	0.049	0.178	0.196	0.217	0.311	0.411	0.493	1.145
SSWC	C	-0.064	-0.029	0.000	0.001	0.010	0.016	0.020	0.149	0.167	0.187	0.282	0.381	0.463	1.116
ASWC	D	-0.064	-0.030	-0.001	0.000	0.009	0.015	0.019	0.148	0.166	0.187	0.281	0.381	0.463	1.115
$C/k^{1/2}$	E	-0.073	-0.039	-0.010	-0.009	0.000	0.006	0.011	0.139	0.157	0.178	0.273	0.372	0.454	1.106
PBM	F	-0.079	-0.045	-0.016	-0.015	-0.006	0.000	0.005	0.133	0.151	0.172	0.267	0.366	0.448	1.100
ASSWC	G	-0.084	-0.049	-0.020	-0.019	-0.011	-0.005	0.000	0.129	0.147	0.167	0.262	0.361	0.443	1.096
Dunn22	H	-0.212	-0.178	-0.149	-0.148	-0.139	-0.133	-0.129	0.000	0.018	0.039	0.133	0.232	0.314	0.967
Dunn23	I	-0.231	-0.196	-0.167	-0.166	-0.157	-0.151	-0.147	-0.018	0.000	0.020	0.115	0.214	0.296	0.949
Dunn62	J	-0.251	-0.217	-0.187	-0.187	-0.178	-0.172	-0.167	-0.039	-0.020	0.000	0.095	0.194	0.276	0.928
DB	K	-0.346	-0.311	-0.282	-0.281	-0.273	-0.267	-0.262	-0.133	-0.115	-0.095	0.000	0.099	0.181	0.834
VRC	L	-0.445	-0.411	-0.381	-0.381	-0.372	-0.366	-0.361	-0.232	-0.214	-0.194	-0.099	0.000	0.082	0.734
Dunn11	M	-0.527	-0.493	-0.463	-0.463	-0.454	-0.448	-0.443	-0.314	-0.296	-0.276	-0.181	-0.082	0.000	0.652
C-Index	N	-1.179	-1.145	-1.116	-1.115	-1.106	-1.100	-1.096	-0.967	-0.949	-0.928	-0.834	-0.734	-0.652	0.000
	Mean	0.944	0.910	0.880	0.880	0.871	0.865	0.860	0.731	0.713	0.693	0.598	0.499	0.417	-0.236

Fig. 13 Mean values (bottom bar) and their differences (cells) for Pearson correlation between relative and external (Jaccard) criteria: $k_{\max} = 50$, $k^* \in \{2, 4, 6\}$, $n \in \{2, 3, 4\}$.



		A	B	C	D	E	F	G	H	I	J	K	L	M	N
Point-Biserial	A	0.000	0.077	0.080	0.085	0.086	0.123	0.146	0.223	0.376	0.426	0.480	0.635	0.653	0.832
SWC	B	-0.077	0.000	0.002	0.008	0.009	0.046	0.069	0.146	0.299	0.349	0.403	0.557	0.575	0.755
$C/k^{1/2}$	C	-0.080	-0.002	0.000	0.005	0.006	0.043	0.067	0.143	0.297	0.346	0.401	0.555	0.573	0.753
ASWC	D	-0.085	-0.008	-0.005	0.000	0.001	0.038	0.061	0.138	0.292	0.341	0.395	0.550	0.568	0.748
PBM	E	-0.086	-0.009	-0.006	-0.001	0.000	0.037	0.060	0.137	0.291	0.340	0.394	0.549	0.567	0.747
ASSWC	F	-0.123	-0.046	-0.043	-0.038	-0.037	0.000	0.023	0.100	0.253	0.303	0.357	0.512	0.530	0.709
SSWC	G	-0.146	-0.069	-0.067	-0.061	-0.060	-0.023	0.000	0.077	0.230	0.280	0.334	0.488	0.507	0.686
VRC	H	-0.223	-0.146	-0.143	-0.138	-0.137	-0.100	-0.077	0.000	0.154	0.203	0.257	0.412	0.430	0.610
Dunn22	I	-0.376	-0.299	-0.297	-0.292	-0.291	-0.253	-0.230	-0.154	0.000	0.049	0.104	0.258	0.276	0.456
Dunn23	J	-0.426	-0.349	-0.346	-0.341	-0.340	-0.303	-0.280	-0.203	-0.049	0.000	0.054	0.209	0.227	0.407
Dunn62	K	-0.480	-0.403	-0.401	-0.395	-0.394	-0.357	-0.334	-0.257	-0.104	-0.054	0.000	0.154	0.173	0.352
Dunn11	L	-0.635	-0.557	-0.555	-0.550	-0.549	-0.512	-0.488	-0.412	-0.258	-0.209	-0.154	0.000	0.018	0.198
DB	M	-0.653	-0.575	-0.573	-0.568	-0.567	-0.530	-0.507	-0.430	-0.276	-0.227	-0.173	-0.018	0.000	0.180
C-Index	N	-0.832	-0.755	-0.753	-0.748	-0.747	-0.709	-0.686	-0.610	-0.456	-0.407	-0.352	-0.198	-0.180	0.000
	Mean	0.986	0.909	0.906	0.901	0.900	0.863	0.840	0.763	0.610	0.560	0.506	0.351	0.333	0.154

Fig. 14 Mean values (bottom bar) and their differences (cells) for Pearson correlation between relative and external (Jaccard) criteria: $k_{\max} = 50$, $k^* \in \{2, 4, 6\}$, $n \in \{22, 23, 24\}$.

results obtained from the overall collection of 972 data sets. Note that, for compactness, Figs. 13–17 present only 3 (the best ones in this part of the experiments) out of those 17 variants of Dunn's index reviewed in Section 2.1.4, namely, Dunn22, Dunn23, and Dunn62.

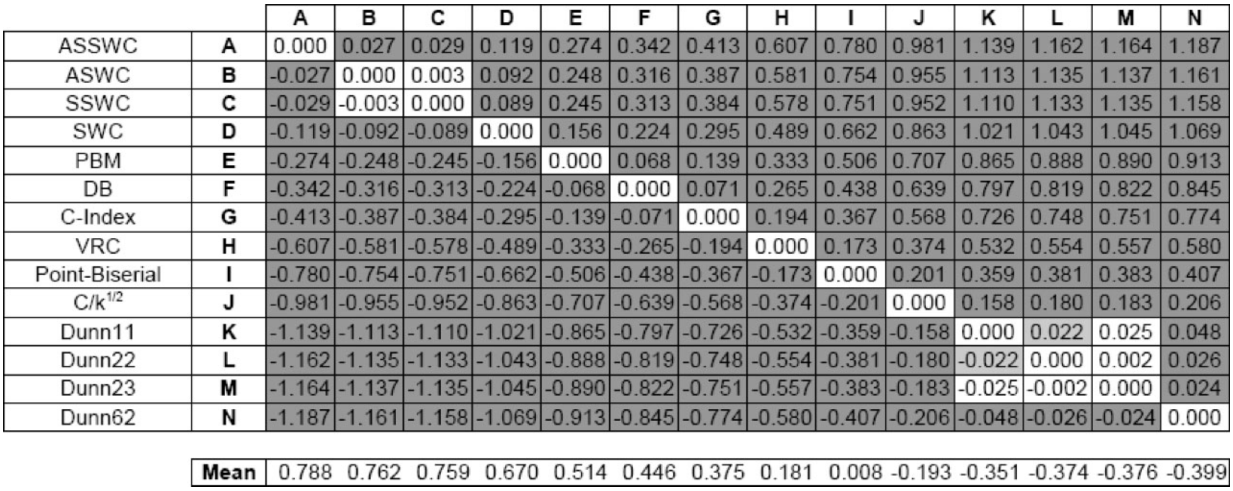
Analyses analogous to those reported in Figs. 13–17 were performed for all possible evaluation scenarios, i.e. by varying the correlation measure (WGK or Pearson), the external criterion (Jaccard or ARI), and the maximum number of clusters ($k_{\max} = 23$ or $k_{\max} = 50$), thus resulting in 40 different tables (2 correlation measures × 2 external criteria × 2 values for $k_{\max}$ × 5 collections of data sets). Such tables are omitted here for compactness reasons and are available at http://www.icmc.usp.br/~campello/Sub_Pages/Selected_Publications.htm.

The main conclusions that can be derived from the 32 tables that correspond to the four subcategories of data sets with different values of n and k^* are the following[10]:

(1) For data sets formed by fewer dimensions, point-biserial generally presented better results in the presence of fewer clusters.

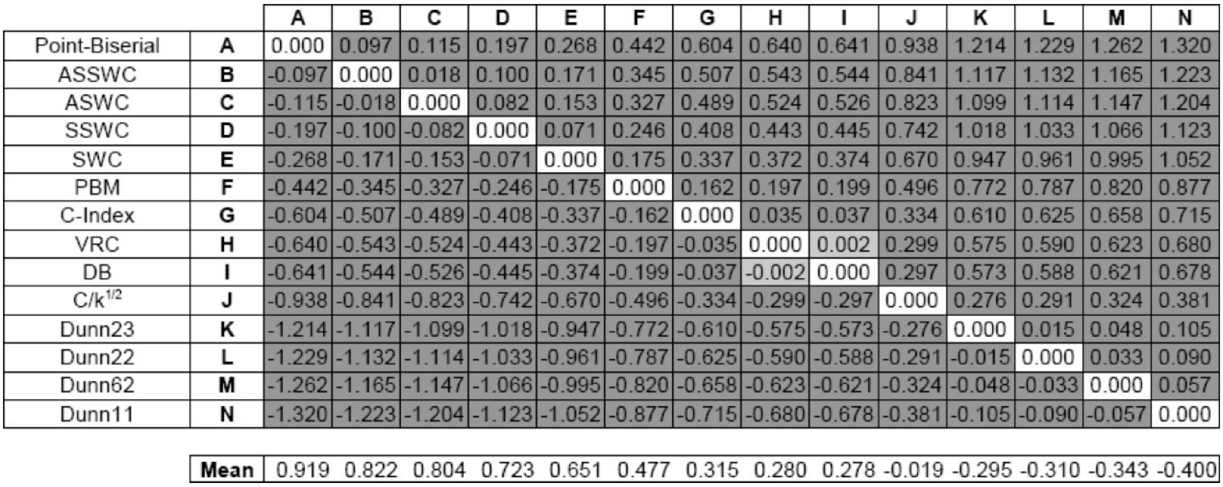
(2) VRC generally presented better results in the presence of fewer clusters for data sets formed by more dimensions. In addition, when data sets with fewer clusters are considered, VRC generally presented

[10] These conclusions refer to the individual performance of each criterion whose behavior varied clearly across the different subcategories of data.



		A	B	C	D	E	F	G	H	I	J	K	L	M	N
ASSWC	A	0.000	0.027	0.029	0.119	0.274	0.342	0.413	0.607	0.780	0.981	1.139	1.162	1.164	1.187
ASWC	B	-0.027	0.000	0.003	0.092	0.248	0.316	0.387	0.581	0.754	0.955	1.113	1.135	1.137	1.161
SSWC	C	-0.029	-0.003	0.000	0.089	0.245	0.313	0.384	0.578	0.751	0.952	1.110	1.133	1.135	1.158
SWC	D	-0.119	-0.092	-0.089	0.000	0.156	0.224	0.295	0.489	0.662	0.863	1.021	1.043	1.045	1.069
PBM	E	-0.274	-0.248	-0.245	-0.156	0.000	0.068	0.139	0.333	0.506	0.707	0.865	0.888	0.890	0.913
DB	F	-0.342	-0.316	-0.313	-0.224	-0.068	0.000	0.071	0.265	0.438	0.639	0.797	0.819	0.822	0.845
C-Index	G	-0.413	-0.387	-0.384	-0.295	-0.139	-0.071	0.000	0.194	0.367	0.568	0.726	0.748	0.751	0.774
VRC	H	-0.607	-0.581	-0.578	-0.489	-0.333	-0.265	-0.194	0.000	0.173	0.374	0.532	0.554	0.557	0.580
Point-Biserial	I	-0.780	-0.754	-0.751	-0.662	-0.506	-0.438	-0.367	-0.173	0.000	0.201	0.359	0.381	0.383	0.407
$C/k^{1/2}$	J	-0.981	-0.955	-0.952	-0.863	-0.707	-0.639	-0.568	-0.374	-0.201	0.000	0.158	0.180	0.183	0.206
Dunn11	K	-1.139	-1.113	-1.110	-1.021	-0.865	-0.797	-0.726	-0.532	-0.359	-0.158	0.000	0.022	0.025	0.048
Dunn22	L	-1.162	-1.135	-1.133	-1.043	-0.888	-0.819	-0.748	-0.554	-0.381	-0.180	-0.022	0.000	0.002	0.026
Dunn23	M	-1.164	-1.137	-1.135	-1.045	-0.890	-0.822	-0.751	-0.557	-0.383	-0.183	-0.025	-0.002	0.000	0.024
Dunn62	N	-1.187	-1.161	-1.158	-1.069	-0.913	-0.845	-0.774	-0.580	-0.407	-0.206	-0.048	-0.026	-0.024	0.000
	Mean	0.788	0.762	0.759	0.670	0.514	0.446	0.375	0.181	0.008	-0.193	-0.351	-0.374	-0.376	-0.399

Fig. 15 Mean values (bottom bar) and their differences (cells) for Pearson correlation between relative and external (Jaccard) criteria: $k_{\max} = 50$, $k^* \in \{12, 14, 16\}$, $n \in \{2, 3, 4\}$.



		A	B	C	D	E	F	G	H	I	J	K	L	M	N
Point-Biserial	A	0.000	0.097	0.115	0.197	0.268	0.442	0.604	0.640	0.641	0.938	1.214	1.229	1.262	1.320
ASSWC	B	-0.097	0.000	0.018	0.100	0.171	0.345	0.507	0.543	0.544	0.841	1.117	1.132	1.165	1.223
ASWC	C	-0.115	-0.018	0.000	0.082	0.153	0.327	0.489	0.524	0.526	0.823	1.099	1.114	1.147	1.204
SSWC	D	-0.197	-0.100	-0.082	0.000	0.071	0.246	0.408	0.443	0.445	0.742	1.018	1.033	1.066	1.123
SWC	E	-0.268	-0.171	-0.153	-0.071	0.000	0.175	0.337	0.372	0.374	0.670	0.947	0.961	0.995	1.052
PBM	F	-0.442	-0.345	-0.327	-0.246	-0.175	0.000	0.162	0.197	0.199	0.496	0.772	0.787	0.820	0.877
C-Index	G	-0.604	-0.507	-0.489	-0.408	-0.337	-0.162	0.000	0.035	0.037	0.334	0.610	0.625	0.658	0.715
VRC	H	-0.640	-0.543	-0.524	-0.443	-0.372	-0.197	-0.035	0.000	0.002	0.299	0.575	0.590	0.623	0.680
DB	I	-0.641	-0.544	-0.526	-0.445	-0.374	-0.199	-0.037	-0.002	0.000	0.297	0.573	0.588	0.621	0.678
$C/k^{1/2}$	J	-0.938	-0.841	-0.823	-0.742	-0.670	-0.496	-0.334	-0.299	-0.297	0.000	0.276	0.291	0.324	0.381
Dunn23	K	-1.214	-1.117	-1.099	-1.018	-0.947	-0.772	-0.610	-0.575	-0.573	-0.276	0.000	0.015	0.048	0.105
Dunn22	L	-1.229	-1.132	-1.114	-1.033	-0.961	-0.787	-0.625	-0.590	-0.588	-0.291	-0.015	0.000	0.033	0.090
Dunn62	M	-1.262	-1.165	-1.147	-1.066	-0.995	-0.820	-0.658	-0.623	-0.621	-0.324	-0.048	-0.033	0.000	0.057
Dunn11	N	-1.320	-1.223	-1.204	-1.123	-1.052	-0.877	-0.715	-0.680	-0.678	-0.381	-0.105	-0.090	-0.057	0.000
	Mean	0.919	0.822	0.804	0.723	0.651	0.477	0.315	0.280	0.278	-0.019	-0.295	-0.310	-0.343	-0.400

Fig. 16 Mean values (bottom bar) and their differences (cells) for Pearson correlation between relative and external (Jaccard) criteria: $k_{\max} = 50$, $k^* \in \{12, 14, 16\}$, $n \in \{22, 23, 24\}$.

superior performance in the presence of more dimensions. Then, it can be asserted that VRC exhibited better results for data sets formed by more dimensions and fewer clusters.

(3) In general, C-index showed better results for data sets that contain more clusters, independently of the number of dimensions. Opposite behavior was observed from PBM, the variants of Dunn, and $C/\sqrt{k}$. In what concerns $C/\sqrt{k}$, this is an expected result. Indeed, it is an outcome of a confounding effect that is inherent to the nature of this criterion, as previously discussed in Section 5.3.

(4) The DB criterion showed a tendency to provide better results for data sets with fewer dimensions. Such a tendency was more prominent for data sets with fewer clusters.

In relative terms, the main conclusions that can be derived from the eight tables corresponding to the assessments involving the complete collection of 972 data sets are the following:

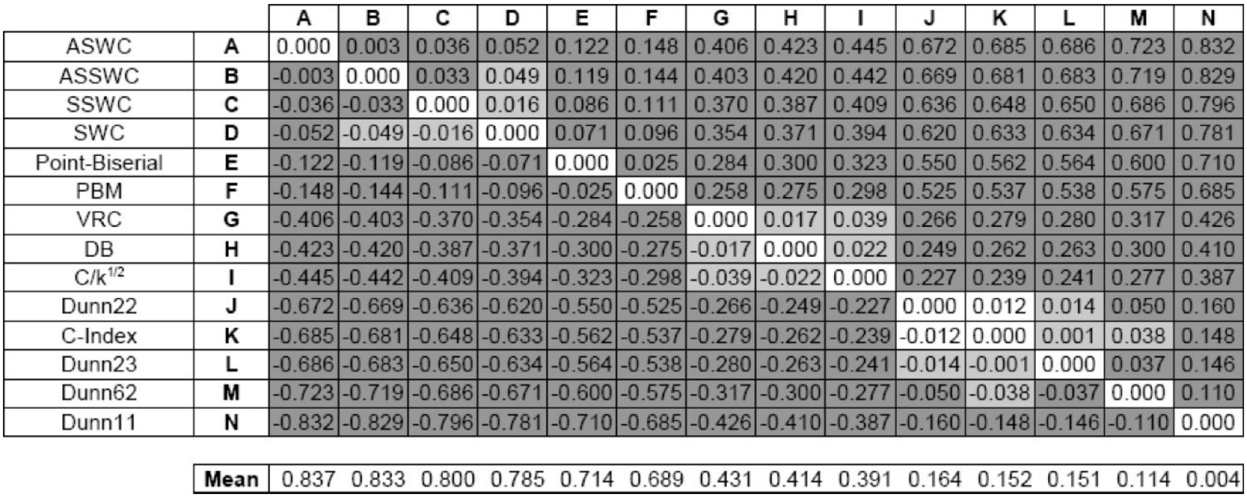
(1) (ASWC, ASSWC, SWC, SSWC, point-biserial, PBM)>VRC>(C-index, DB, Dunn22, Dunn23, Dunn62, Dunn11)

(2) (C-index, DB, Dunn22, Dunn23, Dunn62)>Dunn11

(3) ASWC>SSWC

(4) (SWC, ASWC)>PBM

(5) (DB, Dunn22)>Dunn62



		A	B	C	D	E	F	G	H	I	J	K	L	M	N
ASWC	A	0.000	0.003	0.036	0.052	0.122	0.148	0.406	0.423	0.445	0.672	0.685	0.686	0.723	0.832
ASSWC	B	-0.003	0.000	0.033	0.049	0.119	0.144	0.403	0.420	0.442	0.669	0.681	0.683	0.719	0.829
SSWC	C	-0.036	-0.033	0.000	0.016	0.086	0.111	0.370	0.387	0.409	0.636	0.648	0.650	0.686	0.796
SWC	D	-0.052	-0.049	-0.016	0.000	0.071	0.096	0.354	0.371	0.394	0.620	0.633	0.634	0.671	0.781
Point-Biserial	E	-0.122	-0.119	-0.086	-0.071	0.000	0.025	0.284	0.300	0.323	0.550	0.562	0.564	0.600	0.710
PBM	F	-0.148	-0.144	-0.111	-0.096	-0.025	0.000	0.258	0.275	0.298	0.525	0.537	0.538	0.575	0.685
VRC	G	-0.406	-0.403	-0.370	-0.354	-0.284	-0.258	0.000	0.017	0.039	0.266	0.279	0.280	0.317	0.426
DB	H	-0.423	-0.420	-0.387	-0.371	-0.300	-0.275	-0.017	0.000	0.022	0.249	0.262	0.263	0.300	0.410
$C/k^{1/2}$	I	-0.445	-0.442	-0.409	-0.394	-0.323	-0.298	-0.039	-0.022	0.000	0.227	0.239	0.241	0.277	0.387
Dunn22	J	-0.672	-0.669	-0.636	-0.620	-0.550	-0.525	-0.266	-0.249	-0.227	0.000	0.012	0.014	0.050	0.160
C-Index	K	-0.685	-0.681	-0.648	-0.633	-0.562	-0.537	-0.279	-0.262	-0.239	-0.012	0.000	0.001	0.038	0.148
Dunn23	L	-0.686	-0.683	-0.650	-0.634	-0.564	-0.538	-0.280	-0.263	-0.241	-0.014	-0.001	0.000	0.037	0.146
Dunn62	M	-0.723	-0.719	-0.686	-0.671	-0.600	-0.575	-0.317	-0.300	-0.277	-0.050	-0.038	-0.037	0.000	0.110
Dunn11	N	-0.832	-0.829	-0.796	-0.781	-0.710	-0.685	-0.426	-0.410	-0.387	-0.160	-0.148	-0.146	-0.110	0.000

Mean	0.837	0.833	0.800	0.785	0.714	0.689	0.431	0.414	0.391	0.164	0.152	0.151	0.114	0.004

Fig. 17 Mean values (bottom bar) and their differences (cells) for Pearson correlation between relative and external (Jaccard) criteria: $k_{\max} = 50$, $k^* \in \{2, 4, 6, 12, 14, 16\}$, $n \in \{2, 3, 4, 22, 23, 24\}$.

where ">" means that a significant difference has been observed between the respective criteria with respect to at least one of the statistical tests adopted (W/M-W and/or Friedman). A more conservative reading of the results, in which better performance is asserted if and only if there are significant differences with respect to both tests, allows concluding that:

(1) (SWC, SSWC, ASWC, ASSWC, point-biserial, PBM) >(VRC, C-index, DB, Dunn22, Dunn23, Dunn62, Dunn11)

(2) VRC>(Dunn22, Dunn23, Dunn62, Dunn11)

(3) (C-index, DB)>Dunn11

(4) (SWC, ASWC)>PBM

(5) DB>Dunn62

To summarize, one can conclude from this second part of the experiments that the silhouettes and point-biserial presented the best results in relative terms, followed by PBM and VRC. Among these criteria, only the silhouettes showed to be almost insensitive (more robust) throughout all the assessed scenarios.

7. CONCLUSIONS AND PERSPECTIVES

This paper has presented an overview of 40 relative clustering validity criteria that includes an original comparative asymptotic analysis of their computational complexities. This overview has been divided into two parts, one of them dedicated to optimization-like criteria and the other one devoted to difference-like criteria. An effective formulation to convert difference-like criteria into optimization-like counterparts has been discussed.

An alternative, possibly complementary methodology for comparing relative clustering validity criteria has been described. Such a methodology has been especially designed to get around conceptual limitations that may take place when using the comparison paradigm traditionally adopted in the literature. In particular: (i) it does not rely on the assumption that the accuracy of a validity criterion can be precisely quantified by the relative frequency with which it indicates as the best partition a partition with the right number of clusters; (ii) it does not rely on the assumption that a mistake made by a certain validity criterion when assessing a set of candidate partitions of a given data set can be quantified by the absolute difference between the right (known) number of clusters in those data and the number of clusters contained in the partition elected as the best one; and (iii) it does not rely solely on the correctness of the single partition elected as the best one according to that criterion. Getting rid of such over-simplified assumptions may make the alternative comparison methodology more suitable to assess the performances of validity criteria when trying to distinguish between better and worse partitions embedded in difficult application scenarios. In spite of this, the alternative methodology should be seen as complementary to the traditional one, especially because: (i) it is applicable to the comparison of optimization-like criteria only. When comparing simple stopping rules, which are only able to estimate the number of clusters in data, the traditional methodology is possibly the only choice; and (ii) the traditional methodology focuses on a specific aspect of the problem (the number of clusters). Insightful results—such as those highlighted in the end of Section 5.3—can be revealed based on the simultaneous application of both methodologies.

An extensive experimental comparison of the performances of those 40 relative clustering validity criteria surveyed in the manuscript has been accomplished. The experiments have involved a collection of 962,928 partitions derived by running five different clustering algorithms (four hierarchical ones and k-means) over 1080 different data sets of a given class of interest. Stringent analyses of the results—based on two different external validity indexes, two different correlation measures, two different intervals for the maximum acceptable number of clusters, and two different statistical tests—yielded a number of interesting conclusions about the behavior of the criteria under investigation. The behavior of the criteria when assessing partitions of data sets with different amounts of clusters and dimensions (attributes) has been explored. In addition, experiments performed by Milligan and Cooper in their 1985 paper [14] have been faithfully reproduced with the inclusion of several validity criteria that were not covered in that classic study.

Aside from a number of detailed observations related to particular evaluation scenarios that have been discussed along the manuscript, there are some elucidative overall conclusions that have been derived from all the experiments and the corresponding results. In general, the best performances have been associated with the silhouettes, PBM, VRC, and point-biserial. Among them, the silhouettes have apparently shown the most robust performances with respect to the different evaluation scenarios. The others have exhibited some sort of sensitivity to at least one aspect of the analysis (e.g. the number of attributes and/or clusters in the data sets), mainly point-biserial and VRC. Inasmuch as the simplified silhouettes have exhibited global performances that are comparable to that of the original criterion, they might be preferred when dealing with very large data sets (due to their lighter computational requirements).

As a word of caution, it is worth remarking that the above results and conclusions hold for a particular collection of data sets. Since such a collection is reasonably representative of a particular class, namely, data with volumetric clusters following normal distributions, it seems legitimate to believe that similar results are likely to be observed for other data sets of this class. However, nothing can be presumed about data sets that do not fall within this class, at least not before new experiments involving such data are performed.

Finally, it is also worth remarking that all the discussions related to the experimental portion of this work make sense under the assumption that the adopted reference partitions with known clusters following normal distributions satisfy the user's expectations of *right partition*. This seems to be particularly meaningful for unsupervised clustering practitioners looking for volumetric, well-behaved clusters.

Studies involving different classes of data sets and the comparison of clustering validity criteria of a different nature (e.g. for evaluation of fuzzy partitions) are interesting subjects that deserve further research.

ACKNOWLEDGMENTS

This work was supported by the Brazilian National Research Council (CNPq) and the Research Foundation of the State of São Paulo (Fapesp).

REFERENCES

[1] L. Kaufman and P. Rousseeuw, Finding Groups in Data, New York, NY, USA, Wiley, 1990.

[2] B. S. Everitt, S. Landau, and M. Leese, Cluster Analysis, (4th ed.), London, UK, Edward Arnold, 2001.

[3] G. Gan, C. Ma, and J. Wu, Data Clustering: Theory, Algorithms, and Applications, Philadelphia, PA, USA, ASA-SIAM, 2007.

[4] A. K. Jain, M. N. Murty, and P. J. Flynn, Data clustering: a review, ACM Comput Surv 31 (1999), 264–323.

[5] R. Xu and D. C. Wunsch II. Survey of clustering algorithms, IEEE Trans Neural Netw 16 (2005), 645–678.

[6] L. Wang and X. Fu, Data Mining with Computational Intelligence, Secaucus, NJ, USA, Springer, 2005.

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

[8] J. A. Hartigan, Clustering Algorithms, New York, Wiley, 1975.

[9] R. Xu and D. C. Wunsch II. Clustering, New York, NY, USA, Wiley/IEEE Press, 2009.

[10] J. B. McQueen. Some methods of classification and analysis of multivariate observations, In Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, California, USA, 1967, 281–297.

[11] A. Dempster, N. Laird, and D. Rubin, Maximum likelihood from incomplete data via the em algorithm, J R Stat Soc 39(1) (1977), 1–38.

[12] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Sparks, NV, USA, Springer, 2001.

[13] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, On clustering validation techniques, J Intell Inf Syst 17 (2001), 107–145.

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

[15] D. L. Davies and D. W. Bouldin, A cluster separation measure, IEEE Trans Pattern Anal Mach Intell 1 (1979), 224–227.

[16] R. B. Calinski and J. Harabasz, A dentrite method for cluster analysis, Commun Stat 3 (1974), 1–27.

[17] J. C. Dunn, Well separated clusters and optimal fuzzy partitions, J Cybern 4 (1974), 95–104.

[18] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, Clustering validity checking methods: Part II, SIGMOD Rec 31 (2002), 19–27.

[19] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J Comput Appl Math 20 (1987), 53–65.

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

[21] J. C. Bezdek and N. R. Pal, Some new indexes of cluster validity, IEEE Trans Syst Man Cybern B 28(3) (1998), 301–315.

[22] U. Maulik and S. Bandyopadhyay, Performance evaluation of some clustering algorithms and validity indices, IEEE Trans Pattern Anal Mach Intell 24(12) (2002), 1650–1654.

[23] E. H. Ruspini, Numerical methods for fuzzy clustering, Inf Sci 2 (1970), 319–350.

[24] E. R. Hruschka, R. J. G. B. Campello, A. A. Freitas, and A. C. P. L. F. de Carvalho, A survey of evolutionary algorithms for clustering, IEEE Trans Syst Man Cybern C 39 (2009), 133–155.

[25] L. Hubert and P. Arabie, Comparing partitions, J Classif 2 (1985), 193–218.

[26] G. Casella and R. L. Berger, Statistical Inference (2th ed.), California, USA, Duxbury Press, 2001.

[27] G. W. Milligan, A monte carlo study of thirty internal criterion measures for cluster analysis, Psychometrika 46(2) (1981), 187–199.

[28] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka, On the comparison of relative clustering validity criteria, In SIAM International Conference on Data Mining, Sparks, NV, USA, 2009, 733–744.

[29] E. R. Hruschka, R. J. G. B. Campello, and L. N. de Castro, Evolving clusters in gene-expression data, Inf Sci 176 (2006), 1898–1927.

[30] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recognit Soc 37 (2004), 487–501.

[31] L. J. Hubert and J. R. Levin, A general statistical framework for assessing categorical clustering in free recall, Psychol Bull 10 (1976), 1072–1080.

[32] F. B. Baker and L. J. Hubert, Measuring the power of hierarchical clustering analysis, J Am Stat Assoc 70(349) (1975), 31–38.

[33] F. J. Rohlf, Methods of comparing classifications, Ann Rev Ecol Syst 5 (1974), 101–113.

[34] M. G. Kendall and J. D. Gibbons, Rank Correlation Methods, London, UK, Edward Arnold, 1990.

[35] R. J. G. B. Campello and E. R. Hruschka, On comparing two sequences of numbers and its applications to clustering analysis, Inf Sci 179 (2009), 1025–1039.

[36] D. A. Ratkowsky and G. N. Lance, A criterion for determining the number of groups in a classification, Aust Comput J 10 (1978), 115–117.

[37] R. S. Hill, A stopping rule for partitioning dendrograms, Bot Gaz 141 (1980), 321–324.

[38] H. P. Friedman and J. Rubin, On some invariant criteria for grouping data, J Am Stat Assoc 62 (1967), 1159–1178.

[39] A. J. Scott and M. J. Symons, Clustering methods bases on likelihood ratio criteria, Biometrics 27 (1971), 387–397.

[40] S. J. Arnold, A test for clusters, J Mark Res 19 (1979), 545–551.

[41] F. H. C. Marriot, Practical problems in a method of cluster analysis, Biometrics 27 (1971), 501–514.

[42] G. H. Ball and D. J. Hall, Isodata, A Novel Method of Data Analysis and Pattern Classification. Menlo Park, Stanford Research Institute, NTIS No. AD 699616, 1965.

[43] J. O. McClain and V. R. Rao, CLUSTISZ: a program to test for the quality of clustering of a set of objects, J Mark Res 12 (1975), 456–460.

[44] I. J. Good, An index of separateness of clusters and a permutation test for its statistical significance, J Stat Comput Simul 15 (1982), 81–84.

[45] W. M. Rand, Objective criteria for the evaluation of clustering methods, J Am Stat Assoc 66 (1971), 846–850.

[46] G. Saporta and G. Youness. Comparing two partitions: some proposals and experiments, In Proceedings in Computational Statistics, W. Hardle, ed., Berlin, Germany, Physica-Verlag, 2002.

[47] L. Denoeud, H. Garreta, and A. Guenoche. Comparison of distance indices between partitions, In Proceedings 11th Conference of the Applied Stochastic Models and Data Analysis (ASMDA), Brest, France, 2005.

[48] R. Sharan and R. Shamir, Click: a clustering algorithm with applications to gene expression analysis, In Proceedings 8th International Conference on Intelligent Systems for Molecular Biology, Vienna, Austria, 2000, 307–316.

[49] D. Jiang, C. Tang, and A. Zhang, Cluster analysis for gene-expression data: a survey, IEEE Trans Knowl Data Eng 16 (2004), 1370–1386.

[50] E. B. Fowlkes and C. L. Mallows, A method for comparing two hierarchical clustering, J Am Stat Assoc 46 (1983), 553–569.

[51] D. L. Wallace, Comment on "a method for comparing two hierarchical clustering", J Am Stat Assoc 78 (1983), 569–576.

[52] G. W. Milligan and M. C. Cooper, A study of the comparability of external criteria for hierarchical cluster analysis, Multivariate Behav Res 21 (1986), 441–458.

[53] L. A. Goodman and W. H. Kruskal, Measures of association for cross-classifications, J Am Stat Assoc 49 (1954), 732–764.

[54] M. F. Triola, Elementary Statistics, Don Mills, Ontario, Canada, Addison Wesley, 1999.

[55] X. Wu, V. Kumar, J. R. Wuinlan, J. Ghosh, K. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, Top 10 algorithms in data mining, Knowl Inf Syst 14 (2008), 1–37.