ELSEVIER

Pattern Recognition Letters
journal homepage: www.elsevier.com

Fast Computation of Cluster Validity Measures for Bregman Divergences and Benefits

Marco Capó[a,**], Aritz Pérez[a], Jose A. Lozano[a,b]

[a] *Basque Center of Applied Mathematics, Bilbao 48009, Spain*
[b] *Intelligent Systems Group, Department of Computer Science and Artifitial Intelligence, University of the Basque Country UPV/EHU, San Sebastián 20018, Spain*

Article history:

Partitional Clustering, Number of clusters, Silhouette index, Davies-Bouldin, Caliński-Harabasz, Bregman divergences

ABSTRACT

Partitional clustering is one of the most relevant unsupervised learning and pattern recognition techniques. Unfortunately, one of the main drawbacks of these methodologies refer to the fact that the number of clusters is generally assumed to be known beforehand and automating its selection is not straightforward. On the same token, internal validity measures, such as the Silhouette index, Davies-Bouldin and Caliński-Harabasz measures have emerged as the standard techniques to be used when comparing the goodness of clustering results obtained via different clustering methods. These measures take into consideration both the inter and intra-cluster simmilarities and can be adapted to different metrics. Unfortunately, their used has been hindered due to their large computational complexities, which are commonly quadratic with respect to the number of instances of the data set. In this work, we show that the time complexity of computing the most popular internal validity measures can be utterly reduced by making used of the within-cluster errors and different properties of the Bregman divergences. This contribution ultimately allows us to massively speed-up the selection of an adequate number of clusters for a given data set as verified with extensive empirical comparisons.

1. Introduction

Clustering is an unsupervised machine learning technique used to divide unlabeled data into groups that contain data points that are *similar* to each other and *dissimilar* from those in other clusters (Jain et al., 1999). This is, the partition is done in such a way that intra-cluster similarity is high and the inter-cluster similarity is low. Ultimately, this data analysis technique tries to unveil the inherent structure of a given set of point as well as it allows pattern identification. Clustering is widely used in many fields, such as text mining, image analysis, and bioinformatics (Liu et al., 2010; Capo et al., 2020; Jain et al., 1999).

In an unsupervised learning scenario, one commonly does not possess any prior knowledge of the number of clusters. For this reason, throughout the years multiple validation measures (VM) for clustering results has been designed (Caliński and Harabasz, 1974; Davies and Bouldin, 1979; Hämäläinen et al., 2017). The aim of a VM is to measure the quality of a given clustering based on the compactness and separation of the clusters. A comprehensive review of clustering validation techniques is provided in Hämäläinen et al. (2017). Even though such clustering goodness metrics can be used for the different types of clustering techniques, e.g., partitional, hierarchical, density-based, probabilistic, grid-based, spectral and fuzzy clustering (Aggarwal and Clustering, 2014; Jain et al., 1999; Xie and Beni, 1991). In this work, we focus our analysis on partitional clustering methodologies.

**mcapo@bcamath.com (Marco Capo), aperez@bcamath.com (Aritz Pérez), ja.lozano@ehu.eus (Jose A.Lozano).

1.1. Partitional Clustering

Partitional clustering algorithms group the data into K, commonly non-overlapping, clusters, each of which is represented by a certain prototype. This prototype is commonly selected as the centroid or medoid of the cluster. Given a set of n data points (instances) $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in $\mathbb{R}^d$, an integer K, a dissimilarity measure, $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ and a set of points $D \subseteq \mathbb{R}^d$, partitional clustering commonly attempts to determine a set of points $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_K\} \subseteq D$, so as to minimize the following error function:

$$E^X(P) = \sum_{\mathbf{x} \in X} s(\mathbf{x}, \mathbf{p}_\mathbf{x}), \text{ with } \mathbf{p}_\mathbf{x} = \arg\min_{\mathbf{p} \in P} s(\mathbf{x}, \mathbf{p}) \quad (1)$$

The obtained clustering, $\mathcal{P} = \{P_1, \ldots, P_K\}$, is then given by $P_k = \{\mathbf{x} \in X : k = \arg\min_{i \in \{1,\ldots,K\}} s(\mathbf{x}, \mathbf{p}_i)\}$, for all $k \in \{1, \ldots, K\}$. Unfortunately, finding the solution to this kind of problem is commonly NP-Hard (Aloise et al., 2009; Mahajan et al., 2009). For this reason, partitional clustering is mainly constituted by iterative refinement-based heuristics that approximate their solution (Kaufman and Rousseeuw, 1987).

Depending on the selection of the set $D \subseteq \mathbb{R}^d$ and the dissimilarity $s(\cdot, \cdot)$, we have fairly different problems and algorithms to approximate their solution. In particular, for $D = \mathbb{R}^d$ and $s(\cdot, \cdot)$ any Bregman divergence, we have which is probably the most popular partitional clustering techinique, the K**-means** algorithm (Banerjee et al., 2005; Lloyd, 1982). This heuristic converges to a local minima of Eq.1. In order to do so, this approach iteratively relocates the data points between clusters until a locally optimal partition is attained. The main advantage of the K-means algorithm refers to its low computational complexity, which is $\mathcal{O}(n \cdot K)$. Besides the easiness of its implementation, this method also reduces monotonically the K-means error, i.e., every step of the algorithm generates a more competitive solution to the K-means problem.

If we instead consider $D = X$ in Eq.1, we have the K**-medoids** problem. This problem is originally defined for any dissimilarity measure, $s(\cdot, \cdot)$, and so, K-medoids clustering has emerged as a standard approach in those scenarios where the K-means algorithm can not be applied (Ng and Han, 2002; Schubert and Rousseeuw, 2019). Besides this, for K-medoids clustering, one selects instances from the data set as representatives for each cluster, which are known to be more robust to outliers than other commonly used choices (Han and Kamber, 2011). The Partitioning Around Medoids algorithm (Kaufman and Rousseeuw, 1987), or just **PAM**, is probably the most commonly used heuristic for the K-medoids problem. PAM recursively performs swaps between each pair medoid-non-medoid and then the swap that leads to the largest error decrease is performed until no error reduction can be achieved. The recursive stage of PAM has a $\mathcal{O}(K \cdot (n-K)^2)$ time cost. In order to deal with such a large computational demand, different speed-ups to PAM have been proposed that trade optimality for runtime, out of which both **CLARA** and **CLARANS** stand out as the most popular (Kaufman, 1986; Ng and Han, 2002; Schubert and Rousseeuw, 2019).

CLARA (Kaufman, 1986; Kaufman and Rousseeuw, 2008) applies PAM on multiple subsamples, keeping the best clustering. Commonly the size of the subsample selected is $n' = 40 + 2 \cdot K$ (Schubert and Rousseeuw, 2019). In general, if $n' \in \mathcal{O}(K)$, then the algorithm is $\mathcal{O}(K^3)$. On the other hand, CLARANS (Ng and Han, 2002) works on the entire data set, but only explores a subset of all the possible swaps between medoids and non-medoids. In particular, CLARANS interprets the search space as a high-dimensional hypergraph, where each edge corresponds to the considered swap between a medoid and a non-medoid. Making use of this graph, CLARANS performs a randomnized greedy exploration, where the first edge that reduces the error function is followed until no edge can be found after a predefined number of attempts. Unfortunately, its time complexity is still $\mathcal{O}(n^2)$ (Zhang and Couloigner, 2005), which makes it unsuitable for massive data applications.

1.2. Number of Clusters Selection

As commented in Section 1.1, for partitional clustering algorithms the number of clusters, K, is usually assumed to be known in advanced. Unfortunately, selecting an adequate number of clusters for a given data set is not a simple task. Throughout the years different alternatives have been proposed to approach this issue. In particular, Pelleg et al. (2000) proposed technique, for the K-means problem, that determines a number of clusters by performing recursive bisections of a given set of clusters as long as the obtained Bayesian Information Criterion (BIC) is improved. On the other hand, Hamerly and Elkan (2004) suggested a method to learn K, for the K-means problem as well, based on a statistical test for determining whether instances are a random sample from a Gaussian distribution with arbitrary dimension and covariance matrix. The previous methodologies make assumptions on the underlying distribution of instances belonging to the same cluster and are not commonly used in practice (Hämäläinen et al., 2017).

Among other popular approaches, one can mention the Elbow rule, which consists on evaluating and plotting the error function, Eq.1, which decreases monotonically w.r.t. K, for different values of K and picking the knee of the obtained curve as the number of clusters to use. More importantly, a last family of measures that has gained a lot of popularity, for both selecting the most adequate clustering algorithm for a given data set, as well as for tunning its parameters, are the clustering validation measures (Hämäläinen et al., 2017). In the following section, we describe in detail these metrics and their features.

1.2.1. Clustering Validation Measures

Clustering validation measures evaluate the goodness of clustering results (Maulik and Bandyopadhyay, 2002) and so, they can be used to compare the quality of the results

obtained for different methods, as well as for identifying the optimal values for the clustering parameters, such as the number of clusters, K (Maulik and Bandyopadhyay, 2002; Van Craenendonck and Blockeel, 2015). These measures are divided into two categories: external clustering validation indexes and internal clustering validation indexes. The main difference is whether or not external information is used for clustering validation, as for instance labels. An example of external validation measure is entropy, which evaluates the purity of clusters based on a given class labels (Liu et al., 2010). Unfortunately, in practice, external information such as class labels is often not available in many application scenarios. Therefore, in the situation where there is no external information available, internal validation measures are the only option for cluster validation (Hämäläinen et al., 2017; Liu et al., 2010; Maulik and Bandyopadhyay, 2002).

Internal validation measures are often based on the following two criteria: i) *Compactness*: Its commonly a variance-based metric that quantifies how closely related objects in a cluster are, e.g., maximum/average pairwise distance or center-based distance (Liu et al., 2010), ii) *Separation*: It evaluates how well-separated a cluster is from other clusters, e.g., the pairwise distances between cluster centers or the pairwise minimum distances between objects in different clusters. Internal validation measures are larger when either compactness and/or separation are maximized.

A complete overview and analysis of internal validity measures can be found in (Arbelaitz et al., 2013; Milligan and Cooper, 1985; Van Craenendonck and Blockeel, 2015; Vendramin et al., 2010). In particular, the most popular measures in the literature are the **Silhoutte** index (Rousseeuw, 1987), the **Davies-Bouldin** measure (Davies and Bouldin, 1979) and the **Caliński-Harabasz** measure (Caliński and Harabasz, 1974):

Definition 1. *The Silhoutte index for an instance* $\boldsymbol{x} \in P_k \subseteq X$*, is defined as* $sh(\{\boldsymbol{x}\}) = \frac{b(\boldsymbol{x})-a(\boldsymbol{x})}{\max\{a(\boldsymbol{x}),b(\boldsymbol{x})\}}$*, where* $a(\boldsymbol{x}) = \frac{1}{|P_k|-1} \cdot \sum_{\boldsymbol{y}\in P_k, \boldsymbol{y}\neq\boldsymbol{x}} s(\boldsymbol{y},\boldsymbol{x})$ *and* $b(\boldsymbol{x}) = \min_{i\neq k} \frac{1}{|P_i|} \cdot \sum_{\boldsymbol{y}\in P_i} s(\boldsymbol{y},\boldsymbol{x})$.

It should be noted that $sh(\{\mathbf{x}\}) \in [-1,1]$. Furthermore, observe that $b(\mathbf{x})$ accounts for the cluster separability and $a(\mathbf{x})$ for the cluster compactness. Therefore, ideally $b(\mathbf{x}) \gg a(\mathbf{x})$ for all $\mathbf{x} \in X$, i.e., $sh(\{\mathbf{x}\})$ should be close to 1. The Silhouette value of a partition, $\mathcal{P} = \{P_1, \ldots, P_K\}$, is the average of these values over all instances: $sh(\mathcal{P}) = \frac{1}{n} \cdot \sum_{k=1}^{K} \sum_{\mathbf{x}\in P_k} sh(\{\mathbf{x}\})$ and, in general, its computational complexity is $\mathcal{O}(n^2)$ (Van Craenendonck and Blockeel, 2015).

Definition 2. *The Davies-Bouldin measure, for a partition* $\mathcal{P}$*, is defined as* $db(\mathcal{P}) = \frac{1}{K} \cdot \sum_{k=1}^{K} D_k$*, where* $D_k = \max_{l\neq k} R_{kl}$*,* $R_{kl} = \frac{S_k+S_l}{M_{kl}}$*,* $M_{kl} = s(\boldsymbol{c}_k, \boldsymbol{c}_l)$*,* $S_k = d_k \cdot \sum_{\boldsymbol{x}\in P_k} s(\boldsymbol{x}, \boldsymbol{c}_k)$*,* $\boldsymbol{c}_k = \overline{P_k}$ *and* d_k *a positive constant.*

The Davies-Bouldin measure was originally defined taking the dissimilarity function, $s(\cdot,\cdot)$, to be a l_p-norm and $d_k = (\frac{1}{|P_k|})^{\frac{1}{p}}$, for $p \in \mathbb{N}$ (Davies and Bouldin, 1979). In general, observe that the time complexity of computing the Davies-Bouldin measure is $\mathcal{O}(n \cdot K)$ and, since its numerator measures the compactness of the generated clusters and its denominator considers the separability between them, this measure should be minimized.

Definition 3. *The Caliński-Harabasz measure is defined as* $ch(\mathcal{P}) = \frac{(n-K)\cdot \sum_{k=1}^{K} |P_k|\cdot s(\boldsymbol{c}_k, \overline{X})}{(K-1)\cdot E^X(C)}$*, where* $C = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K\}$ *and* $\boldsymbol{c}_k = \overline{P_k}$.

Note that the time complexity of evaluating the Caliński-Harabasz measure is $\mathcal{O}(n \cdot K)$ and, since the numerator contains the separability information and the denominator measures the compacity, this index should be maximized.

1.3. Contribution

In this work, we assume the dissimilarity $s(\cdot,\cdot)$ to belong to the Bregman divergence family of metrics and make use of its properties to utterly reduce the running times required to compute the most commmonly used internal clustering validation measures, see Section 1.2.1. This will for instance allow us to easily automate the selection of an appropriate number of clusters without exceeding the actual time complexity of the predefined partitional algorithm, which is a problem of high interest in the unsupervised learning community (Hamerly and Elkan, 2004; Pelleg et al., 2000).

At this point it must be highlighted that the Bregman divergences are a broad class of dissimilarity measures indexed by strictly convex functions that are useful in a wide range of areas, among which are statistical learning and data mining, computational geometry, natural sciences, speech processing and information theory (Banerjee et al., 2005; Gray et al., 1980). A Bregman divergence is formally defined as follows:

Definition 4 (Bregman divergence (Bregman, 1967))**.** *Let* $\phi : \Omega \mapsto \mathbb{R}$ *be a strictly convex function defined on a convex set* $\Omega \subseteq \mathbb{R}^d$*, such that* ϕ *is differentiable on* $ri(\Omega)$[1]*, assumed to be non-empty. The Bregman divergence* $s : \Omega \times ri(\Omega) \mapsto \mathbb{R}_+$ *is defined as*

$$s(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - < \boldsymbol{x} - \boldsymbol{y}, \nabla\phi(\boldsymbol{y}) >, \tag{2}$$

In Table 1, we show a few of the well-known dissimilarity measures that belong to this family of metrics:

Table 1: Examples of Bregman divergences.

Divergence	Ω	$\phi(\mathbf{x})$
Kullback-Leibler	d-simplex[2]	$\mathbf{x} \cdot \log \mathbf{x}$
squared Euclidean	$\mathbb{R}^d$	$\|\mathbf{x}\|^2$
Itakura-Saito	$\mathbb{R}^d_{++}$[3]	$-\log \mathbf{x}$

[1] $ri(\Omega)$ stands for the relative interior of the convex set Ω.

The rest of the article is organized as follows: In Section 2, we elaborate on the theoretical aspects of our proposal and extend on the computation of internal validity measures, as well as of the proposed algorithm for selecting the number of clusters. Afterwards, in Section 3, we compare the proposed methodology empirically with respect to the standard approach in terms of quality of the obtained solution and computational load.

2. Efficient Number of Cluster Selection via Internal Validity Measures

In general, when performing partitional clustering, we assume the number of clusters to be predefined (Jain et al., 1999; Jain, 2010). On the other hand, since one of the main features of validity measures, such as the Silhouette index, is that they do not get neccesarily improved by increasing the number of clusters, as it happens when using the error function Eq.1 in the Elbow rule, they are commonly used to automate the selection of the number of clusters (Hämäläinen et al., 2017). Unfortunately, one of the factors that hinders the use of these indexes to perform such a task is related to the added computational load of computing them (Lensen et al., 2017). As discussed in Section 1.2.1, for instance computing the Silhoutte index from scratch is $\mathcal{O}(n^2)$, which is more time consuming than running a typical partitional clustering algorithms such as the K-means algorithm, which is $\mathcal{O}(n \cdot K)$.

In this section, we will exploit different properties of Bregman divergences that have been previously deduced in papers such as Banerjee et al. (2005). These properties have been applied on the clustering context, in particular considering the typical ouput and of partitional clustering algorithms (within-cluster errors for each cluster) to fasten the computation of the most commonly used internal validity measures, i.e., Silhouette index, Davies-Bouldin measure and Caliński-Harabasz measure, in such a way that we can efficiently select an appropriate number of clusters, without increasing the complexity of the clustering method.

It is important to highlight that this sort of approaches can be easily extended to other validity measures and dissimilarities within the Bregman divergences. In this work, we restrict our analysis to the most popular clustering validity measures.

2.1. Efficient Computation of the Silhouette Index, Davies-Bouldin and Caliński-Harabasz Measures

As previously commented in Section 1.1, partitional clustering algorithms usually minimizes error functions of the form of Eq.1. In order to calculate $E^X(P)$, which is needed to evaluate the stopping criteria of the algorithm, the selected partitional clustering method computes the pairwise-distance matrix between the instances in X and the set of prototypes P and uses this information to evaluate $E^X(P) = \sum_{k=1}^{K} E^{P_k}(\{\mathbf{p}_k\})$, where $E^{P_k}(\{\mathbf{p}_k\})$ is the within-cluster error for the k^{th} cluster, for all $k \in \{1, \ldots, K\}$. In the upcoming result, we re-write the internal validity measures presented in Section 1.2.1 in terms of the within-cluster errors for all Bregman divergences:

Theorem 1. *Given a clustering $\mathcal{P} = \{P_1, \ldots, P_K\}$ and let $\boldsymbol{c}_k$ be the centroid of P_k, i.e., $\boldsymbol{c}_k = \overline{P_k}$, for all $k \in \{1, \ldots, K\}$, and $s(\cdot, \cdot)$ a Bregman divergence, then i) $sh(\{\boldsymbol{x}\}) = \frac{b(\boldsymbol{x}) - a(\boldsymbol{x})}{\max\{a(\boldsymbol{x}), b(\boldsymbol{x})\}}$, for $\boldsymbol{x} \in P_k$, where $a(\boldsymbol{x}) = \frac{1}{|P_k|-1} \cdot [E^{P_k}(\{\boldsymbol{c}_k\}) + |P_k| \cdot s(\boldsymbol{c}_k, \boldsymbol{x})]$ and $b(\boldsymbol{x}) = \min_{i \neq k} \frac{1}{|P_i|} \cdot [E^{P_i}(\{\boldsymbol{c}_i\}) + |P_i| \cdot s(\boldsymbol{c}_i, \boldsymbol{x})]$, ii) $db(\mathcal{P}) = \frac{1}{K} \cdot \sum_{k=1}^{K} \max_{l \neq k} \frac{d_k \cdot E^{P_k}(\{\boldsymbol{c}_k\}) + d_l \cdot E^{P_l}(\{\boldsymbol{c}_l\})}{s(\boldsymbol{c}_k, \boldsymbol{c}_l)}$ and iii) $ch(\mathcal{P}) = \frac{(n-K) \cdot [E^X(\{\overline{X}\}) - E^X(C)]}{(K-1) \cdot E^X(C)}$, where $C = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K\}$.*

Proof. First of all observe that, for $\mathbf{c}_k = \overline{P_k}$ and $k \in \{1, \ldots, K\}$ and $\mathbf{x} \in \mathbb{R}^d$, using Proposition 1 in Banerjee et al. (2005), we have

$$\begin{aligned} \sum_{\mathbf{y} \in P_k} s(\mathbf{y}, \mathbf{x}) &= \sum_{\mathbf{y} \in P_k} s(\mathbf{y}, \mathbf{c}_k) + |P_k| \cdot [\phi(\mathbf{c}_k) - \phi(\mathbf{x}) - \\ &\quad - < \mathbf{c}_k - \mathbf{x}, \nabla\phi(\mathbf{x}) >] \\ &= \sum_{\mathbf{y} \in P_k} s(\mathbf{y}, \mathbf{c}_k) + |P_k| \cdot s(\mathbf{c}_k, \mathbf{x}) \end{aligned} \tag{3}$$

Using Eq.3, we can re-write $a(\mathbf{x})$ and $b(\mathbf{x})$, for any $\mathbf{x} \in \mathbb{R}^d$:

$$\begin{aligned} a(\mathbf{x}) &= \frac{1}{|P_k| - 1} \cdot \sum_{\mathbf{y} \in P_k, \mathbf{y} \neq \mathbf{x}} s(\mathbf{y}, \mathbf{x}) = \frac{1}{|P_k| - 1} \cdot \sum_{\mathbf{y} \in P_k} s(\mathbf{y}, \mathbf{x}) \\ &= \frac{1}{|P_k| - 1} \cdot E^{P_k}(\{\mathbf{x}\}) = \frac{1}{|P_k| - 1} \cdot [E^{P_k}(\{\mathbf{c}_k\}) + \\ &\quad + |P_k| \cdot s(\mathbf{c}_k, \mathbf{x})] \end{aligned} \tag{4}$$

and

$$\begin{aligned} b(\mathbf{x}) &= \min_{i \neq k} \frac{1}{|P_i|} \cdot \sum_{\mathbf{y} \in P_i} s(\mathbf{y}, \mathbf{x}) = \min_{i \neq k} \frac{1}{|P_i|} \cdot E^{P_i}(\{\mathbf{x}\}) \\ &= \min_{i \neq k} \frac{1}{|P_i|} \cdot [E^{P_i}(\{\mathbf{c}_i\}) + |P_i| \cdot s(\mathbf{c}_i, \mathbf{x})] \end{aligned} \tag{5}$$

From Eqs.4-5, we have the result for the Silhoutte index. Furthermore, for the Davies-Bouldin measure, we have that $S_k = d_k \cdot E^{P_k}(\{\mathbf{c}_k\})$ and $M_{kl} = s(\mathbf{c}_k, \mathbf{c}_l)$. In words, $R_{kl} = \frac{d_k \cdot E^{P_k}(\{\mathbf{c}_k\}) + d_l \cdot E^{P_l}(\{\mathbf{c}_l\})}{s(\mathbf{c}_k, \mathbf{c}_l)}$ and, therefore,

$$\begin{aligned} db(\mathcal{P}) &= \frac{1}{K} \cdot \sum_{k=1}^{K} D_k = \frac{1}{K} \cdot \sum_{k=1}^{K} \max_{l \neq k} R_{kl} \\ &= \frac{1}{K} \cdot \sum_{k=1}^{K} \max_{l \neq k} \frac{d_k \cdot E^{P_k}(\{\mathbf{c}_k\}) + d_l \cdot E^{P_l}(\{\mathbf{c}_l\})}{s(\mathbf{c}_k, \mathbf{c}_l)} \end{aligned} \tag{6}$$

[2] d-simplex=$\{\mathbf{x} = (x_1, \ldots, x_d) \in [0, 1]^d : x_1 + \ldots + x_d = 1\}$.

[3] $\mathbb{R}^d_{++} = \{\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_i > 0 \ \forall i\}$ and $\mathbb{R}^d_{+} = \{\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_i \geq 0 \ \forall i\}$

Finally, for the Caliński-Harabasz measure, we observe that, using Eq.3, the following equality holds

$$\begin{aligned} E^X(\{\overline{X}\}) &= \sum_{k=1}^{K}\sum_{\mathbf{x}\in P_k} s(\mathbf{x},\overline{X}) = \sum_{k=1}^{K}[E^{P_k}(\{\mathbf{c}_k\}) + \\ &\quad +|P_k|\cdot s(\mathbf{c}_k,\overline{X})] \\ &= E^X(C) + \sum_{k=1}^{K}|P_k|\cdot s(\mathbf{c}_k,\overline{X}) \end{aligned} \tag{7}$$

and so, $ch(\mathcal{P}) = \frac{(n-K)\cdot[E^X(\{\overline{X}\})-E^X(C)]}{(K-1)\cdot E^X(C)}$ □

Theorem 1 shows that, if the dissimilarity considered is a Bregman divergence, we can express the Silhoutte index, Davies-Bouldin and Caliński-Harabasz measures in terms of the within-cluster errors obtained for the given clustering and the pairwise-distance between the cluster centroids (for the Davies-Bouldin measure). Hence, the computational cost required to compute this internal validation measures can be utterly reduced by taking advantage information that can be extracted from the output of the most common partitional clustering algorithms:

Proposition 1. *Given a clustering $\mathcal{P}$, i) $sh(\mathcal{P})$ can be computed in $\mathcal{O}(n\cdot K)$ time, ii) $db(\mathcal{P})$ can be computed in $\mathcal{O}(K^2)$ time if the clustering algorithm is centroid-based for Eq.1, otherwise, if it is medoid-based, in $\mathcal{O}(\max\{n,K^2\})$ time, and iii) $ch(\mathcal{P})$ can be evaluated in $\mathcal{O}(n)$ time, for all Bregman divergences.*

Proof. First of all, we assume the partitional algorithm to be centroid-based for Eq.1. Observe that computing $sh(\{\mathbf{x}\}) = \frac{b(\mathbf{x})-a(\mathbf{x})}{\max\{a(\mathbf{x}),b(\mathbf{x})\}}$, for $\mathbf{x}\in P_k$, where $a(\mathbf{x}) = \frac{1}{|P_k|-1}\cdot[E^{P_k}(\{\mathbf{c}_k\})+|P_k|\cdot s(\mathbf{c}_k,\mathbf{x})]$ and $b(\mathbf{x}) = \min_{i\neq k}\frac{1}{|P_i|}\cdot[E^{P_i}(\{\mathbf{c}_i\})+|P_i|\cdot s(\mathbf{c}_i,\mathbf{x})]$, is $\mathcal{O}(K)$. Hence, $sh(\mathcal{P}) = \frac{1}{n}\cdot\sum_{\mathbf{x}\in X} sh(\{\mathbf{x}\})$ can be calculated in $\mathcal{O}(n\cdot K)$. On the same token, evaluating $db(\mathcal{P}) = \frac{1}{K}\cdot\sum_{k=1}^{K}\max_{l\neq k}\frac{d_k\cdot E^{P_k}(\{\mathbf{c}_k\})+d_l\cdot E^{P_l}(\{\mathbf{c}_l\})}{s(\mathbf{c}_k,\mathbf{c}_l)}$ is just $\mathcal{O}(K^2)$ due to the pairwise distance computations between centroids. On the other hand, $ch(\mathcal{P}) = \frac{(n-K)\cdot[E^X(\{\overline{X}\})-E^X(C)]}{(K-1)\cdot E^X(C)}$ is $\mathcal{O}(n)$ due to the computation of $E^X(\{\overline{X}\})$.

If the clustering algorithm is medoid-based, then the within-cluster errors provided by the algorithm are of the form $E^{P_k}(\{\mathbf{m}_k\})$, for some $\mathbf{m}_k\in P_k$, for all $k\in\{1,\ldots,K\}$. Since $s(\cdot,\cdot)$ is assumed to be a Bregman divergence, then $E^{P_k}(\{\mathbf{c}_k\}) = E^{P_k}(\{\mathbf{m}_k\})-|P_k|\cdot s(\mathbf{c}_k,\mathbf{m}_k)$ by Eq.3. For this reason, the added cost of computing the internal validation measure is due to the computation of $\mathbf{c}_k = \overline{P_k}$, for all $k\in\{1,\ldots,K\}$, which is $\mathcal{O}(n)$, and evaluating $s(\mathbf{c}_k,\mathbf{m}_k)$, for all $k\in\{1,\ldots,K\}$, which is $\mathcal{O}(K)$. Therefore, the additional cost is $\mathcal{O}(n)$ for the medoid-based case and so, evaluating $db(\mathcal{P})$, in this setting, is $\mathcal{O}(\max\{n,K^2\})$. □

In Proposition 1, we verify that by re-using the within-cluster errors, we can reduce the computational complexity of the Silhoutte index from $\mathcal{O}(n^2)$ to $\mathcal{O}(n\cdot K)$, the Davies-Bouldin measure decreases from $\mathcal{O}(n\cdot K)$ to $\mathcal{O}(\max\{n,K^2\})$, in the worst case, and the Caliński-Harabasz measure from $\mathcal{O}(n\cdot K)$ to $\mathcal{O}(n)$. In words, we can evaluate the most commonly used internal validity measures (Van Craenendonck and Blockeel, 2015) in, at most, $\mathcal{O}(n\cdot K)$ time. This is, we can quantify the goodness of the obtained clusterings without exceeding the linear complexity of some of the most popular clustering approach, such as the K-means algorithm.

More importantly, observe that Proposition 1 also eases the use of such measures to determine an adequate number of clusters for a given dataset and clustering algorithm: Given partitional clustering algorithm, *clustering*, e.g., the K-means algorithm, we can recursively applied it for an increasing/decreasing number of clusters. Afterwards, using the obtained clustering, an internal validity measure, *quality*, e.g., the Silhouette index, Davies-Bouldin or Caliński-Harabasz measure, can be evaluated[4]. This process can be repeated until a certain stopping criterion is satisfied: For instance, i) we can run this approach until reaching a predefined maximum/minimum number of clusters and keep the number of clusters that leads to the best *quality* index, ii) we can run the algorithm as long as the *quality* is improved or until after a certain amount of iterations the *quality* index is no longer improved. Regardless of the stopping criterion selected, the complexity of each iteration of this approach is dominated by the complexity of *clustering*.

3. Experiments

In this section, we analyze the performance of the methodology proposed in Section 2, using the time-complexity reduction of the most relevant validity measures, to select a number of clusters for a certain data set and clustering algorithm. In particular, given a clustering algorithm (**Cl**), and a validity measure (**VM**), we consider as baseline running Cl for $K\in\{2,\ldots,50\}$ and keeping the number of clusters that leads to the best VM[5](**Baseline_Cl**). In addition, we consider running Cl for an increasing number of clusters K, starting with $K=2$, and stopping if, after s modifications of the number of clusters, where $s\in\{1,2,5,10\}$, the validity measure, computed via Theorem 1, is not improved (**Cl_s**). Lastly, we also run Cl, for all $K\in\{2,\ldots,50\}$ and compute VM, as in the previous case, via Theorem 1 (**Cl_all**). In terms of the partitional methods considered, in this section we use: i) K-means algorithm (**KM**), ii) Partitioning Around Medoids algorithm (**PAM**) and iii) Clustering Large Applications method (**CLARA**). In terms of the validity measures, we consider i) the Silhouette index (**SH**) and ii) the Davies-Bouldin (**DB**) and iii) Caliński-Harabasz (**CH**) measures, and additionally we compute running times (**RT**). The Bregman

[4] Alternatively, one could evaluate different validity measures for the clustering obtained, for each value K, and select the optimal number of clusters from the corresponding Pareto optimal front. However, as we will see in Section3, the optimal number of clusters tends to be similar among different validity measures.

[5] In this case, VM is evaluated using the `Scikit-Learn` versions of the given index.

divergence used in this section is the squared Euclidean metric, in `https://github.com/MarcoVCapo/VM` results using other dissimilarities can be found.

So as to facilitate the comparison between all considered methodologies, DB, CH and RT are normalized as follows for a given repetition of the experimental setting: i) relative $\mathrm{DB}(M) = \frac{\max_{M' \in \mathcal{M}} \mathrm{DB}(M')}{\mathrm{DB}(M)}$, ii) relative $\mathrm{CH}(M) = \frac{\mathrm{CH}(M)}{\min_{M' \in \mathcal{M}} \mathrm{CH}(M')}$ and iii) relative $\mathrm{RT}(M) = \frac{\mathrm{RT}(M)}{\min_{M' \in \mathcal{M}} \mathrm{RT}(M')}$, where M is an algorithm included in the set of considered methods $\mathcal{M}$. Moreover, SH is not normalized since, regardless of the experimental setting, the index always satisfies $-1 \leq \mathrm{SH}(M) \leq 1$. In this sense, observe that SH, relative CH and relative DB must be maximized and relative RT should be minimized. We have analyzed the performance of the different methods on a wide variety of real data sets obtained from the UCI repository, see Table 2[6]. Given the random nature of the algorithms each experiment has been repeated five times.

Table 2: Information of the data sets.

Data Set	n	d
KEGG	53.413	23
HTRU	17.898	9
Hour	17.379	19
Avila	10.436	11
Gas Turbine	7.384	11

Data Set	n	d
Anuran	7.195	22
Turkiye	5.820	33
Waveform	5.000	41
Gesture	1.444	32
Tripadvisor	980	10

In Figs.1-4, we present the obtained results in terms of SH, relative DB, relative CH and relative RT obtained when maximizing VM following the previosuly mentioned methodologies. We must also point out that, for PAM, we were only able to obtain results for the two smallest data sets (Tripadvisor and Gesture, see Table 2) for the limit running time of 24 hours for a repetition of the experiment.

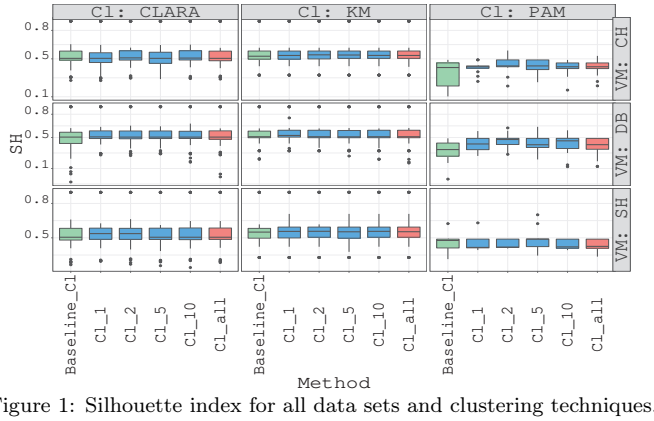

Figure 1: Silhouette index for all data sets and clustering techniques.

[6]Original data sets were normalized via `MinMaxScaler` function of `Scikit-Learn`

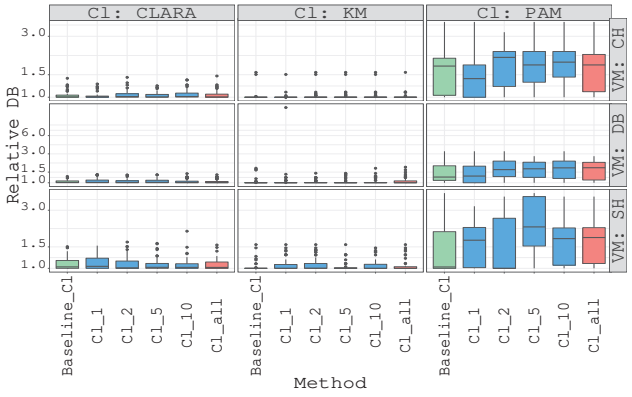

Figure 2: Relative Davies-Bouldin measure for all data sets and clustering techniques.

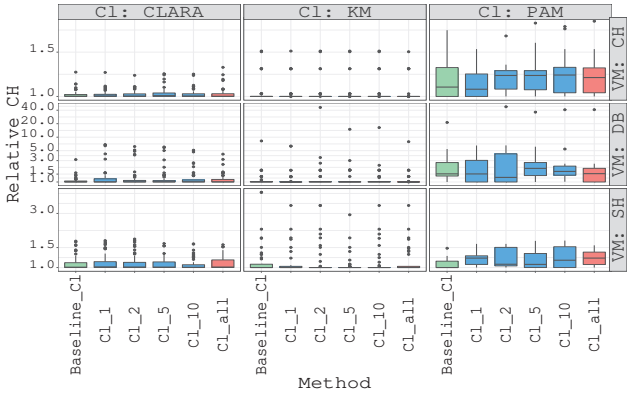

Figure 3: Relative Caliński-Harabasz measure for all data sets and clustering techniques.

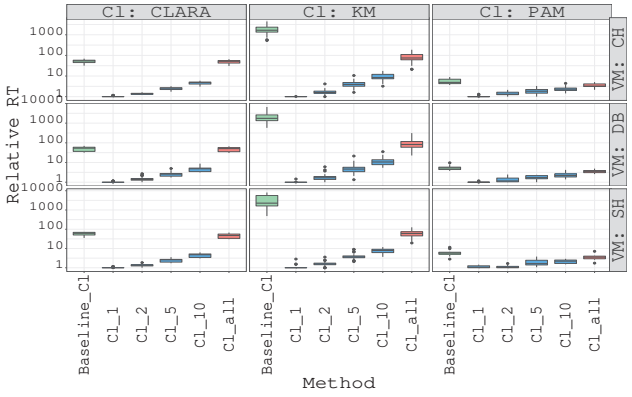

Figure 4: Relative running times for all data sets and clustering techniques.

At first glance, we observe that regardless of the VM index being maximized, the clustering obtained by all techniques considered have similar SH, DB and CH values. In any case, if we focus on the SH values obtained across the different methods, the variability of the obtained results is smaller when VM is SH. Furthermore, as expected, the SH index of Baseline_Cl is of the same order of that of Cl_all for all experimental settings. More importantly, for all settings, Cl_1 already achieved a fairly similar and, sometimes better, SH in comparison to Baseline_Cl, e.g., for KM, the average SH value of Baseline_Cl is 0.55, while, for Cl_1, this value actually increases to 0.56. What's more, if we increase the number of iterations allowed without

improvement, this value increases rapidly, e.g., for Cl_10, we have an average SH value of 0.58.

The results obtained for both relative DB and relative CH follow the same behavior as those obtained and, previously described, for SH. The most interesting variation occurs when Cl is PAM where the most variability on the index values is observed. In any case, the performance of the proposed methodology was still competitve in this setting. For instance, the relative DB obtained by Baseline_Cl is 1.48 for PAM (with VM=DB), while Cl_1 reached and average DB of 1.50. In the case of PAM we do not neccesarily observe an improvement as the number of clusters modifications is increased, which must be linked to the dependence of PAM to the initial set of medoids provided (in this case selected uniformly at random from the data set).

Finally, in terms of the relative RT, we observe the main benefit of our proposal. In particular for both KM and CLARA, we verify a staggering reduction of running time with respect to the baseline. An example of this can be seen in the case of KM and the validity measure SH, where the average relative RT of Baseline_Cl ascends to 3207.86, while, for Cl_1, it is just 1.05, i.e., Baseline_Cl is 3055.10 times slower than Cl_1 and reaches the same clustering. For PAM, such a reduction is less significant due to the computational complexity of the clustering algorithm itself, which is $\mathcal{O}(K \cdot (n-K)^2)$. Still, in this case, Cl_1 was, on average, 5.62 times faster than Baseline_Cl.

Conclusions

In this work, we provide a tool for efficiently selecting the number of clusters for all partitional clustering algorithms. In particular, we demostrate that, for all Bregman divergences, the Silhouette index and the Davies-Bouldin and the Caliński-Harabasz measures can be computed in, at most, $\mathcal{O}(n \cdot K)$ time using the within-cluster errors. Eventhough, empirically we have restricted our analysis to the K-means and K-medoids problem, the use of these results can be easily extended to all partitional clustering techniques, e.g., OPTICS and DBSCAN. This contribution is relevant since it drastically fastens the computation of such validity measures to select an adequate number of clusters. For instance, for the K-means algorithm reductions of up to 3 orders of magnitude in average can be observe with respect to the common brute-force selection of number of clusters. This sort of computations can be further extended to other family of metrics and validity measures.

References

Aggarwal, C.C., Clustering, C.R.D., 2014. Algorithms and applications.

Aloise, D., Deshpande, A., Hansen, P., Popat, P., 2009. Np-hardness of Euclidean sum-of-squares clustering. Machine Learning 75, 245–248.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I., 2013. An extensive comparative study of cluster validity indices. Pattern Recognition 46, 243–256.

Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J., 2005. Clustering with bregman divergences. Journal of machine learning research 6, 1705–1749.

Bregman, L.M., 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR computational mathematics and mathematical physics 7, 200–217.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 3, 1–27.

Capo, M., Perez, A., Lozano, J.A.A., 2020. An efficient split-merge restart for the k-means algorithm. IEEE Transactions on Knowledge and Data Engineering , 1–1.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence , 224–227.

Gray, R., Buzo, A., Gray, A., Matsuyama, Y., 1980. Distortion measures for speech processing. IEEE Transactions on Acoustics, Speech, and Signal Processing 28, 367–376.

Hämäläinen, J., Jauhiainen, S., Kärkkäinen, T., 2017. Comparison of internal clustering validation indices for prototype-based clustering. Algorithms 10, 105.

Hamerly, G., Elkan, C., 2004. Learning the k in k-means. Advances in neural information processing systems 16, 281–288.

Han, J., Kamber, M., 2011. Data mining: Concepts and techniques. Morgan Kaufman Series of Data Management Systems .

Jain, A.K., 2010. Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31, 651–666.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM computing surveys 31, 264–323.

Kaufman, L., 1986. Clustering large data sets. Pattern Recognition in Practice , 425–437.

Kaufman, L., Rousseeuw, P.J., 1987. Clustering by means of medoids. statistical data analysis based on the l1 norm. Y. Dodge, Ed , 405–416.

Kaufman, L., Rousseeuw, P.J., 2008. Clustering large applications (program clara). Finding groups in data: an introduction to cluster analysis , 126–163.

Lensen, A., Xue, B., Zhang, M., 2017. Using particle swarm optimisation and the silhouette metric to estimate the number of clusters, select features, and perform clustering, in: European Conference on the Applications of Evolutionary Computation, Springer. pp. 538–554.

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., 2010. Understanding of internal clustering validation measures, in: 2010 IEEE international conference on data mining, IEEE. pp. 911–916.

Lloyd, S., 1982. Least squares quantization in pcm. IEEE Transactions on Information Theory 28, 129–137.

Mahajan, M., Nimbhorkar, P., Varadarajan, K., 2009. The planar k-means problem is np-hard, in: International Workshop on Algorithms and Computation, pp. 274–285.

Maulik, U., Bandyopadhyay, S., 2002. Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on pattern analysis and machine intelligence 24, 1650–1654.

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

Ng, R.T., Han, J., 2002. Clarans: A method for clustering objects for spatial data mining. IEEE transactions on knowledge and data engineering 14, 1003–1016.

Pelleg, D., Moore, A.W., et al., 2000. X-means: Extending k-means with efficient estimation of the number of clusters., in: Icml, pp. 727–734.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20, 53–65.

Schubert, E., Rousseeuw, P.J., 2019. Faster k-medoids clustering: improving the pam, clara, and clarans algorithms, in: International Conference on Similarity Search and Applications, Springer. pp. 171–187.

Van Craenendonck, T., Blockeel, H., 2015. Using internal validity measures to compare clustering algorithms. Benelearn 2015 Poster

presentations (online) , 1–8.
Vendramin, L., Campello, R.J., Hruschka, E.R., 2010. Relative clustering validity criteria: A comparative overview. Statistical analysis and data mining: the ASA data science journal 3, 209–235.
Xie, X.L., Beni, G., 1991. A validity measure for fuzzy clustering. IEEE Transactions on pattern analysis and machine intelligence 13, 841–847.
Zhang, Q., Couloigner, I., 2005. A new and efficient k-medoid algorithm for spatial clustering, in: International conference on computational science and its applications, Springer. pp. 181–189.