A Combination Approach to Cluster Validation Based on Statistical Quantiles

Amparo Albalate
Institute of Information Technology
University of Ulm
Ulm, Germany
amparo.albalate@uni-ulm.de

David Suendermann
SpeechCycle Labs
New York, USA
david@speechcycle.com

***Abstract*—In this paper, we analyse different techniques to detect the number of clusters in a dataset, also know as *cluster validation techniques*. We also propose a new algorithm based on the combination of several validation indexes to simultaneously validate several partitions of a dataset generated by different clustering techniques and object distances. The existing validation techniques as well as the combination algorithm have been tested on three data sets: a synthesized mixture of Gaussians data set, the NCI60 microarray data set, and the Iris data set. Evaluation results have shown the adequate performance of the proposed approach, even if the input validity scores fail to discover the true number of clusters.**

***Keywords*-cluster validation; quantile;**

I. Introduction

Cluster analysis organises data by abstracting underlying structure either as a grouping of individuals or as a hierarchy of groups [1].

Commonly associated to the usage of cluster algorithms is the problem of estimating the number of classes existing in a dataset. Most clustering algorithms are parameterized approaches, with the target number of clusters k as the most frequent input parameter.

The question about the correct number of clusters in a dataset is not only a topic of recent investigation: Already in the 1970s, with the appearance of the classical clustering approaches, researchers like J. A. Hartigan (k-means) showed strong consciousness about this problem and proposed some metrics for automatically determining the value of k. The general approach is to evaluate the quality of each k-cluster solution provided by the clustering algorithm and select the value of k that originates the optimum partition according to a quality criterion. This particular field of cluster analysis is commonly known as "cluster validation" or "cluster validity". Over the past decades, many approaches for cluster validation have been proposed in parallel to the advances in clustering techniques, such as the Krzanowski and Lai test, the Davies Bouldin index, Silhouette, and more recently, the Gap statistic. Many of them try to minimise/maximise the intra- or inter-cluster dispersion.

Unfortunately, the performance of validation techniques usually depends on the data set or the cluster algorithm used to partition the data. In addition, the distance metric applied prior to clustering has proven a relevant factor for the final cluster solution and may also influence the cluster validity success to determine the optimum number of clusters. In a few cases, prior assumptions about the data set can be adopted which enable the choice of the best fitting clustering technique and distance model. However, unsupervised models are often applied to more complex, multi-dimensional datasets for which little or no prior assumptions can be made.

In this paper, we propose a validity combination strategy to predict the number of clusters in a data set without adopting any prior assumptions about the clustering technique or distance measure. Our approach to cluster validation is to perform multiple simulations on a dataset varying the distance and clustering technique as well as the number of clusters k. Then, the different partitions obtained from these simulations are evaluated in parallel by several cluster validation criteria. A validation redundancy is thereby achieved which can be exploited to measure the agreement/consistency of the different scores at each value k. The new technique described in this paper is based on the calculation of quantile statistics of the validation curves, as explained in the following sections.

The structure of this paper is as follows: In Section II, we give an overview about validation indexes commonly used in the literature. In Section III, we describe the new combination approach. In Sections IV and V, the experimental corpora and results are described. Finally, we draw conclusions in Section VI.

II. Cluster Validation Methods

As outlined in Section I, the determination of the number of clusters in a data set is a principal problem associated with many clustering algorithms.

In the following, we denote $\mathcal{C} = \{C_1, \cdots, C_k\}$, a cluster partition composed of k clusters, and N, the total number of objects in a data set. The cluster validation indexes applied in our experiments are the following:

A. Hartigan

This metric was proposed by J. A. Hartigan for detecting the optimum number of clusters k to be applied in the k-means clustering algorithm [2]:

$$H(k) = \gamma(k)\frac{W(k) - W(k+1)}{W(k+1)}, \quad \gamma(k) = N - k - 1 \quad (1)$$

denoting $W(k)$ the intra-cluster dispersion, defined as the total sum of square distances of the objects to their cluster centroids. The parameter γ is introduced in order to avoid an increasing monotony with increasing k. In this work, we use a small modification to the Hartigan metric, by treating the parameter $W(k)$ as the average intra-cluster distance.

According to Hartigan, the optimum number of clusters is the smallest k which produces $H(k) \leq \eta$ (typically $\eta = 10$). However, in order to allow a better alignment of the Hartigan index to other scores in the combination approach, we have introduced a correction of the index: $Hc(k) = H(k-1)$ and considered a modification of the optimum criterion by maximising $Hc(k)$. In other words, the new criterion maximises the relative improvement at k with respect to $k-1$, in terms of decreasing dispersion. This allows for a direct application of the corrected index $Hc(k)$ in the combination approach without resorting to a previous inversion of the scores.

B. Davies Bouldin Index

The Davies Bouldin index [3] was proposed to find compact and well separated clusters. It is formulated as:

$$DB(k) = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \quad (2)$$

where $\Delta(C_i)$ denotes the intra-cluster distance, calculated as the average distance of all the cluster objects C_i to the cluster medoid, whereas $\delta(C_i, C_j)$ denotes the distance between the clusters C_i and C_j (distance between the cluster medoids). The optimum number of clusters corresponds to the minimum value of $DB(k)$.

C. Krzanowski and Lai Index

This metric belongs to the so-called "elbow models" [4]. These approaches plot a certain quality function over all possible values for k and detect the optimum as the point where the plotted curves reach an elbow, i.e. the value from which the curve considerably decreases or increases. The Krzanowski and Lai index is defined as:

$$KL(k) = \left|\frac{\text{Diff}_k}{\text{Diff}_{k+1}}\right| \quad (3)$$

$$\text{Diff}_k = (k-1)^{\frac{2}{m}}W_{k-1} - k^{\frac{2}{m}}W_k \quad (4)$$

The parameter m represents the feature dimensionality of the input objects (number of attributes), and W_k is calculated as the within-group dispersion matrix of the clustered data:

$$W_k = \sum_{i=1}^{k}\sum_{x_j \in C_i}(x_j - c_i)(x_j - c_i)^T \quad (5)$$

In this case, x_j represents an object assigned to the j^{th} cluster, and c_i denotes the centroid or medoid of the i^{th} cluster. The optimum k corresponds to the maximum of $KL(k)$.

D. Silhouette

This method is based on the *silhouette width*, an indicator for the quality of each object i [5]. The silhouette width is defined as:

$$sil(x_i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (6)$$

where $a(i)$ denotes the average distance of the object i to all objects of the same cluster, and $b(i)$ is the average distance of the object i to the objects of the closest cluster.

Based on the object silhouettes, one can extend the silhouette scores to validate each individual cluster using the average of the cluster object silhouettes:

$$sil(C_j) = \frac{1}{|C_j|}\sum_{x_i \in C_j} sil(x_i) \quad (7)$$

Finally, the silhouette score which validates the whole partition of the data is obtained by averaging the cluster silhouette widths:

$$sil(k) = \frac{1}{k}\sum_{r=1}^{k} sil(C_r) \quad (8)$$

The optimum k maximises $sil(k)$.

E. Gap Statistic

The idea behind the Gap statistic is to compare the validation results of the given data set to an appropriate reference data set drawn from an a-priory distribution [6]. Thereby, this formulation avoids the increasing or decreasing monotony of other validation scores with increasing number of clusters.

First, the intra-cluster distance is averaged over the k clusters:

$$W_k = \sum_{r=1}^{k}\frac{1}{2n_r}\sum_{i,j \in C_r} D(i,j) \quad (9)$$

where n_r denotes the number of elements of the cluster r. The Gap statistic is defined as:

$$Gap(k) = E(log(W_k)) - log(W_k) \quad (10)$$

where $E(log(W_k))$ is the expected logarithm of the average intra-cluster distance. In practice, this expectation

is computed through a Monte-Carlo simulation on a number of sample realizations of a uniform distribution B[1].

$$Gap(k) = \frac{1}{B}\sum_{b}(log(W_{kb})) - log(W_k) \qquad (11)$$

where W_{kb} denotes the average intra-cluster distance of the b^{th} realization of the reference distribution using k clusters. The optimum number of clusters is the smallest value k such that $Gap(k) \geq Gap(k+1) - s_{k+1}$, where s_k is a factor that takes into account the standard deviation of the Monte-Carlo replicates W_{kb}.

III. Combination approach based on quantiles

The proposed approach is based on a combination of validation results using the validity indexes from Section II. First, multiple clustering partitions of a data set have been generated varying the clustering method and distance functions used to cluster the data objects.

- Clustering techniques: In this work, we used four clustering algorithms: the partitioning around medoids (pam) algorithm [7], and the hierarchical complete, centroid and average linkage methods [8].
- Distance functions: The aforementioned algorithms have been applied to two different distance metrics representing the dissimilarity between dataset objects. In this work, we used Euclidean and cosine distances, respectively.

The different clusterings of the dataset have been in turn evaluated using the validity indexes[2] from Section II. In the following, we refer to the validation outcome obtained with each triple (clustering, distance, validation index) as "validation curve". Note that Davies Bouldin scores have been inverted before applying the combination approach so that the optimum can be generalized to the maximum scores and that the Gap statistic has been modified as:

$$Gap'(k) = Gap(k) - Gap(k+1) + s_{k+1} \qquad (12)$$

The proposed method is based on the observation that, although the validation curves may fail to determine the optimum k as a global or local maximum, the correct k is consistently located among the top scores in most cases. This fact motivated the combination of validation scores based on p quantiles.

The p quantile of a random variable X is defined as such value x which is only exceeded by a proportion $1-p$ of the variable samples [9]. In mathematical terms, if denoting the probability density function of the random variable X as $pdf(X)$, the p quantile can be defined as:

$$Q(X,p) = x : \qquad \int_{-\infty}^{x} pdf(X) = p \qquad (13)$$

Figure 1 illustrates this concept for a hypothetic random variable with a normal distribution.

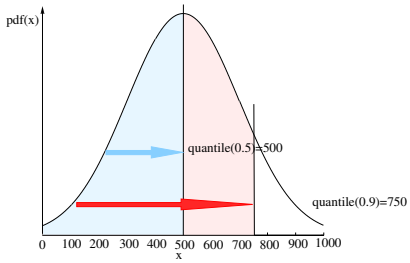

Figure 1. Illustrative example of a p quantile: 0.5 and 0.9 quantiles of variable samples with normal distribution.

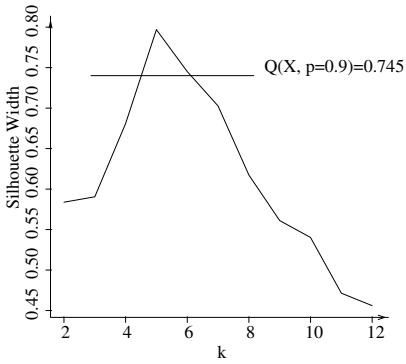

Figure 2. Application of quantiles to a validity curve (Silhouette index, hierarchical average and Euclidean distance). "Top scores" can be identified as such scores that exceed the p quantile level. In this example, $p = 0.90$.

For the application of quantiles to the detection of the number of clusters, the different validation curves are treated as if drawn from a certain probabilistic process. The quantile function is then applied to each single curve, denoted V_i. The p quantile $Q(V_i, p)$ returns the validation score V_{i_p} only exceeded by the $1-p$ proportion of k values in the considered range. This fact is exemplified in Figure 2 for the validation curve obtained by applying the Silhouette index to validate the partitions of the mixture of Gaussians data set using the hierarchical average linkage algorithm and Euclidean distance.

A basic approach to measuring the consensus of validity scores is to directly apply p quantiles to the set of validation curves, $\mathcal{V}$, and counting the number of times that each value k outperforms the score $Q(V_i, p)$. We call this method *Quantile Validation* ($Qvalid(\mathcal{V}, p)$).

However, the *Qvalid* results show a certain dependency on the quantile probability parameter p. On the one hand, using low p values often leads to maximum scores at the

[1]Note that the reference data drawn from this uniform distribution consists of a number N of objects identical to the data set, with identical number of features m. The values of each feature in each object are assigned randomly in the original feature range.

[2]Note: while Hartigan, Krzanowski and Lai, Davies Bouldin, and Silhouette have been used in combination with the four clustering algorithms, the Gap statistic has been only applied to the pam and average linkage algorithms.

Algorithm 1 Quantile validation: $Qvalid(\mathcal{V}, p)$

Input $\mathcal{V}$: set of validation curves, p: quantile parameter
for k=2 to k_{max} **do**
  Qvalid[k] = 0
  for all V_i **do**
    if $V_i[k] \geq$ quantile(p, V_i) **then**
      Qvalid[k] $\leftarrow$Qvalid[k] + 1
    end if
  end for
end for

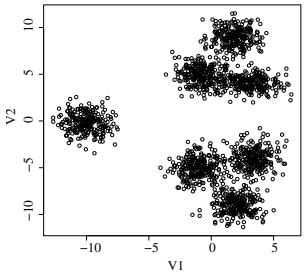

Figure 3. Mixture of seven Gaussians data set

optimum k_{opt}, but these maxima are not unique, given the high proportion of samples which often exceed the levels $Q(V_i, p)$ for low ps. On the other hand, if a high p value is selected, a maximum peak can be clearly discerned, but it is often misplaced to $k \neq k_{opt}$. This happens, in particular, if an increasing or decreasing monotony with k is observed in some validity outcomes. These monotonic effects may be captured in this *Qvalid* result in the form of maximum peaks at low or high ks.

For these reasons, we propose a "supra-consensus" function wich aims at combining the set of quantile validation results obtained with different p values. The algorithm, called *quantile detection*, is performed in three steps: First, the quantile validation is applied to the input validation curves with nine different p values: $p \in [0.1, 0.2, \ldots, 0.9]$. The *Qvalid* results are then modified by casting out (setting to 0) a set ks whose scores are identified as irrelevant according to the information of 0.9 quantiles. Finally, the number of maxima at each k across the modified *Qvalid* scores is returned.

IV. Experimental Corpora

The validity indexes from Section II as well as the combination approach have been evaluated on a synthetic dataset (mixture of Gaussians) and two real data sets.

Mixture of seven Gaussians: The first data set (Figure 3) is a mixture of seven Gaussians in two dimensions. A cluster hierarchy can be observed in the data plot, with three well differentiated and seven less separable clusters.

NCI60 data set: The second data set used in our experiments in the NCI60 dataset [10], publicly available at

Algorithm 2 Quantile detection: $Qdetect(\mathcal{V})$

Input $\mathcal{V}$: set of validation curves
for p=0.1 to 0.9 **do**
  $QD_p[2, \ldots, k_{max}]$ = Qvalid(V, p)
  $QD_p[1]$ = 0
  $QD_p[k_{max} + 1]$ = 0
end for
for all k, p **do**
  if $QD_{0.9}[k] < 0.5 \cdot \max_{k'}(QD_{0.9}[k'])$ **then**
    $QD_p[k]$ = 0
  end if
end for
for k=2 to k_{max} **do**
  Qdetect[k] = 0
  for all p **do**
    if $QD_p[k] = \max_{k'}(QD_p[k'])$ or
    $QD_p[k-1] < QD_p[k] > QD_p[k+1]$ **then**
      Qdetect[k] $\leftarrow$Qdetect[k] + 1
    end if
  end for
end for

[11]. It consists of gene expression data for 60 cell lines derived from different organs and tissues. The data is a 1375x60 matrix where each row represents a gene and each column a cell line related to a human tumour. There are 9 known tumour types: leukemia, colon, breast, prostate, lung, ovarian, renal, CNS, and melanoma, and one unknown.

Iris data set: Finally, we have tested the algorithms on the Iris data set [12], available at the UCI machine learning repository [13]. The data set contains 150 instances with four attributes related to three classes of iris plants (Iris Setosa, Iris Versiculor, Iris Virginica). Two of the classes are linearly separable while one of them is not linearly separable from the other two. Also, a z-score normalisation has been performed on each Iris attribute, $\mathbf{a}$, by using the mean, $\mu(\mathbf{a})$ and standard deviation, $\sigma(\mathbf{a})$: $\bar{\mathbf{a}} = (\mathbf{a} - \nu(\mathbf{a}))/(\sigma(\mathbf{a}))$.

V. Results

This section reports on the validation scores obtained with the validity indexes described in Section II and the combination approach. We have used a maximum number of clusters $k_{max} = 40$.

Validation results obtained on the mixture of Gaussians, NCI60, and Iris data sets are shown in Tables 1, 2 and 3, respectively. The first rows in these tables show an excerpt of the validation curves obtained by applying the validation indexes (Section II) to different partitions of the data set: a) using the partitioning around medoids (pam) with cosine distance, b) pam algorithm with Euclidean distance, c) hierarchical average linkage with cosine distance, and d) hierarchical average with Euclidean distance. The second

Table I

VALIDATION ON THE SEVEN GAUSSIANS DATA SET. The first 20 rows show the validation results obtained with the corrected Hartigan, Krzanowski and Lai, Davies Bouldin, Silhouette and Gap statistics, using the pam and hierarchical average linkage with cosine and Euclidean distance. The next 9 rows show the results of the $Qvalid$ function for $[p = 0.1, 0.2, \ldots, 0.9]$. The second last row shows the counts of maxima across the previous 9 rows. Finally, the last row shows the combined results obtained with the $Qdetect$ algorithm.

Validation Index	Clustering, Distance	k 2	3	4	5	6	7	8	9	10
$Hc(k)$	pam, cos	2422.4**	197.4	4441.9*	1085.3	567.1	556.9	1687.4	524.8	335.4
	pam, Euc	-1459.1	1010.5	185.6	406.6	265.7	53.1	554.6	87.0	37.6
	havg, cos	2707.1	3222.8*	581.0	1376.6***	33.8	976.9	22.7	1670.8**	9.9
	havg, Euc	-1529.3	1897.2*	184.1	388.6***	346.3	526.9**	4.9	3.5	4.9
$KL(k)$	pam, cos	1.154	0.703	4.510	0.861	1.092	0.535	4.004	0.542	3.451
	pam, Euc	0.212	13.773**	0.694	0.748	3.175	0.207	12.875***	0.493	1.262
	havg, cos	2.281	4.396***	0.534	0.496	5.781**	0.460	0.756	0.864	1.350
	havg, Euc	0.412	52.829*	0.336	0.842	0.770	2.667	0.976	1.054	0.989
$DB(k)$	pam, cos	0.896	0.908	0.708*	1.671	2.028	1.508	0.880**	1.327	1.615
	pam, Euc	0.875	0.484*	0.809	0.855	0.667**	0.859	0.791***	0.945	1.123
	havg, cos	0.717	0.445*	0.654	0.784	0.810	0.791**	0.852	0.944	0.982
	havg, Euc	0.728	0.484*	0.669	0.768	0.684	0.602	0.593**	0.608	0.609
$Sil(k)$	pam, cos	0.707	0.590	0.769*	0.707	0.717	0.737	0.738**	0.720	0.702
	pam, Euc	0.530	0.684*	0.546	0.487	0.550***	0.538	0.571**	0.524	0.474
	havg, cos	0.751	0.872*	0.793	0.742	0.668	0.706***	0.689	0.716**	0.677
	havg, Euc	0.437	0.684*	0.584	0.519	0.570	0.610**	0.583	0.545	0.512
$Gap'(k)$	pam, cos	0.578*	-0.713	-0.059	0.204**	0.193	-0.258	-0.101	0.033	0.038
	pam, Euc	-0.275	0.057**	-0.015	-0.010	0.106*	-0.195	0.042***	0.041	0.049
	havg, cos	-0.404	0.257***	0.074	0.478*	-0.036	0.336**	-0.401	0.260	0.283
	havg, Euc	-0.548	0.148*	-0.056	-0.075	-0.151	0.083**	0.082	0.066	0.080***
$Qvalid(p = 0.1)$		29	32	35	32	35	35	33	35	36
$Qvalid(p = 0.2)$		25	32	29	28	32	32	33	33	33
$Qvalid(p = 0.3)$		23	31	27	26	28	29	27	28	28
$Qvalid(p = 0.4)$		19	30	25	23	25	28	22	26	24
$Qvalid(p = 0.5)$		19	30	24	21	25	27	22	23	20
$Qvalid(p = 0.6)$		18	28	23	20	23	26	21	21	18
$Qvalid(p = 0.7)$		17	27	22	20	22	24	18	21	16
$Qvalid(p = 0.8)$		11	27	21	17	18	22	16	20	8
$Qvalid(p = 0.9)$		9	21	13	10	12	13	8	6	3
$Qvalid$ max. counts		0	8*	1	0	0	7**	1	5	2
$Qdetect$		0	8**	1	0	2	9*	0	0	0

last row shows the counts of maxima across the set of *Qvalid* results for $p = [0.1, 0.2, \cdots, 0.9]$, as explained in Section III. The last row shows scores obtained with the *Qdetect* algorithm. For brevity, we only show an excerpt of validation scores at some relevant k values from the analysed range ($k = [2, 40]$). In order to retain significant information despite the omitted results, relevant maxima (minima in the case of Davies Bouldin scores) considering the full analysed range have been marked using asterisk symbols: (*) stands for first maxima, versus (***) for third maxima[3]. Also, the column corresponding to the correct number of clusters has been highlighted in grey background color.

[3]We refer to the global maxima of the validity function as first maxima, while the second and third local maxima are refered to as the second/third maxima. Note that a local maximum is located at k if the score is higher than the values of $k + 1$ and $k - 1$. For edge values ($k = 2, k = 40$), a local maximum is placed if these scores are greater than their adjacent in-range neighbors' scores.

A. Results on the mixture of seven Gaussians data set

Validation results on the mixture of seven Gaussians data set are shown in Table 1. The *Qvalid* results for $p = [0.1, 0.2, \cdots, 0.9]$ are also detailed in Table 1. Note that some columns in grey fonts indicate the set of k values identified as irrelevant and set to zero by the *Qdetect* algorithm, as explained in Section III.

As already discussed in Section IV, the mixture of seven Gaussians data set comprises a hierarchy of three well separable clusters and seven less separable clusters. As can be observed, many validation curves are able to detect the three top clusters but fail to detect the seven Gaussians. In some cases, validation scores reflect the hierarchy by placing first maxima at $k = 3$ and other maxima at $k = 7$ (Hartigan, Silhouette, and Gap statistic using hierarchical average clustering and Euclidean distance; Davies Bouldin with hierarchical average clustering and cosine distance). The Silhouette index produces a third maximum at $k = 7$

Table II
VALIDATION RESULTS ON THE NCI60 DATA SET.

Validation Index	Clustering, Distance	k = 7	8	9	10
$Hc(k)$	pam, cos	1.22	2.18	2.15	0.32
	pam, Euc	-0.21	0.75	0.92***	0.13
	havg, cos	1.84	1.81	0.67	0.09
	havg, Euc	0.15	0.01	2.99*	0.06
$Sil(k)$	pam, cos	0.191	0.205	0.215	0.216
	pam, euc	0.091	0.093	0.104	0.110
	havg, cos	0.134	0.158	0.173	0.172
	havg, Euc	0.092	0.092	0.135**	0.134
$Gap'(k)$	pam, cos	-3.5e-02	-3.5e-02	-6.4e-04	-2.8e-02
	pam, Euc	-0.010	-0.014	0.001	0.0007
	havg, cos	-0.027	-0.004	0.006**	-0.113
	havg, Euc	0.004**	-0.051	0.002	-0.005
$Qvalid$ max. counts		0	0	9*	0
$Qdetect$		0	0	9*	0

Table III
VALIDATION RESULTS ON THE IRIS DATA SET.

Validation Index	Clustering, Distance	k = 2	3	4	5
$Hc(k)$	pam, cos	184.397*	143.755	65.769	9.514
	pam, Euc	-53.932	39.640*	13.550	12.844
	havg, cos	175.052*	143.247	4.289	74.159**
	havg, Euc	-53.932	6.979	1.017	27.146**
$KL(k)$	pam, cos	67.827*	0.067	5.409	1.322
	pam, Euc	5.030	4.090	0.714	5.853***
	havg, cos	4.863***	2.952	13.340*	0.026
	havg, Euc	11.051	0.691	0.513	14.283
$DB(k)$	pam, cos	0.621	0.878	1.230	1.189
	pam, Euc	0.672	0.973	1.133	1.104
	havg, cos	0.617*	0.902	0.911	1.080
	havg, Euc	0.672	0.625	0.547*	0.680
$Sil(k)$	pam, cos	0.737*	0.669	0.646	0.559
	pam, Euc	0.581*	0.447	0.386	0.335
	havg, cos	0.729	0.721	0.662	0.641
	havg, Euc	0.581*	0.480	0.406	0.374
$Gap'(k)$	pam, cos	-0.420	-0.149	0.097**	0.025
	pam, Euc	-0.121	-0.004	0.003	0.059*
	havg, cos	-0.430	0.101**	-0.1643	0.0008
	havg, Euc	0.047	0.109*	-0.093	0.063
$Qvalid$ max. counts		2	7**	1	6
$Qdetect$		3	8*	1	0

if hierarchical average clustering is applied in combination with cosine distance.

The counts of maxima across *Qvalid* also reflects the cluster hierarchy although this strategy results in a first maximum at $k = 3$. A second maximum is obtained at $k = 7$, matching the best validation curves. Note, however, that this maximum is not unique, but other candidates are found at $k = 18$, $k = 23$, and $k = 32$.

Unlike the aforementioned results, the *Qdetect* algorithm detects the correct number of Gaussians as a first, unique maximum at $k = 7$. $k = 3$ is assigned a second maximum.

B. Results on the NCI60 data set

Table II shows validation results on the NCI60 data set. Note that validation curves obtained with the Davies Bouldin and Krzanowski and Lai metrics have not been included. The reason are missing values in the data set.

On this data set, only the (corrected) Hartigan index is able to detect the correct number of classes (9) in one of the validation curves. This occurs when hierarchical average clustering is used in combination with the Euclidean distance. Three other validation curves place local maxima at $k = 9$: second maxima are found at $k = 9$ by Silhouette and Gap statistics with hierarchical average clustering, and a third maximum is placed by Hartigan in combination with the pam algorithm and Euclidean distance.

Regarding our combination approach, both strategies determine the correct number of classes at $k = 9$.

C. Results on the Iris data set

Table III shows the validation results on the Iris data set. As explained in Section IV, this data set is composed of two linearly separable iris types and a third class not linearly separable from the other two. Therefore, most validation curves fail to detect the number of clusters. Validation maxima are often misplaced to $k = 2$. Only Hartigan and Gap statistic, in combination with Euclidean distance, and pam and hierarchical average algorithms, respectively, are able to identify the correct number of classes. Also the Gap statistic with hierarchical average clustering and cosine distance places a second maximum at $k = 3$.

These validation outcomes have been slightly improved by counting the number of maxima across the *Qvalid* results (a second maximum is placed at $k_{opt} = 3$). Finally, as it happened with the previous data sets, the *Qdetect* algorithm is able to identify the correct number of clusters on the Iris data set.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have analysed different existing approaches for discovering the number of clusters in a dataset. In particular, the Hartigan index, Davies Bouldin, Krzanowski and Lai test, the Silhouette width and the Gap statistic have been applied to three data sets: a mixture of seven Gaussians, the NCI60 cancer data set, and the Iris data set.

Motivated by the hypothesis that the validation results may depend on clustering technique and distance model, we have proposed a new algorithm, called *Qdetect*. Our approach is based on a combination of multiple validity outcomes using quantiles.

Experimental results have evidenced that the combined solution (*Qdetect* algorithm) is able to identify the correct number of clusters although many of the individual validation techniques on all three databases fail.

Future work is to investigate the robustness of *Qdetect* on other databases (particularly with larger k_{opt}).

REFERENCES

[1] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, USA: Prentice Hall, 1988.

[2] J. Hartigan, *Clustering Algorithms*. New York, USA: Wiley, 1975.

[3] D. Davies and D. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 224–227, 1979.

[4] W. Krzanowski and Y. Lai, "A criterion for determining the number of groups in a data set using sum of squares clustering," *Biometrics*, vol. 44, pp. 23–34, 1985.

[5] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Jornal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[6] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a dataset via the gap statistic," *Journal of the royal statistical socitety*, vol. 63, pp. 411–423, 2001.

[7] L. Kaufmann and P. Rousseeuw, *Finding Groups in Data. An Introduction to Cluster Analysis*. New York, USA: Wiley, 1990.

[8] B. Everitt, *Cluster Analysis*. London, UK: Heinemann Educ., 1974.

[9] R. Serfling, *Approximation Theorems of Mathematical Statistics*. New York: Wiley, 1980.

[10] D. Ross, U. Scherf, M. Eisen, C. Perou, C. Rees, P. Spellman, V. Iyer, S. Jeffrey, M. van den Rijn, M. Waltham, A. Pergamenschikov, J. Lee, D. Lashkari, D. Shalon, T. Myers, J. Weinstein, D. Botstein, and P. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics*, vol. 24, pp. 227–235, 2000.

[11] `http://genome-www.stanford.eud/nci60`.

[12] R. Fisher, "The use of multiple measurements in taxonomic problems," in *Annual Eugenics*, vol. 7, 1936, pp. 179–188.

[13] `http://archive.ics.uci.edu/ml/datasets/Iris`.