ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Comparison of Cluster Validity Indices and Decision Rules for Different Degrees of Cluster Separation

Sara Kaczyńska[*], Rebecca Marion[*] and Rainer von Sachs

Université catholique de Louvain - ISBA, LIDAM
Voie du Roman Pays 20, 1348 Louvain-la-Neuve - Belgium

Abstract. Clustering algorithms are powerful tools for data exploration but often require the a priori choice of the number of clusters. In practice, cluster validity indices (CVIs) are used to quantify the clustering structure of candidate partitions, then decision rules are applied to the indices to choose the best number of clusters. This study analyzes how dimensionality and the degree of cluster separation impact the choice of the number of clusters according to 7 different indices and various decision rules. In contrast to previous studies, the degree of cluster separation is controlled by a single parameter and several decision rules are tested for each CVI.

1 Introduction

Clustering algorithms partition data points into groups, where similar points are placed in the same group and dissimilar points are placed in different groups. The vast majority of unsupervised clustering algorithms require the a priori choice of the number of clusters, denoted here as the parameter K.

The parameter K can be chosen using *cluster validity indices (CVIs)*. CVIs measure the quality of partitions for different K values, often by quantifying cluster compactness and separation. Then, the number of clusters that enables the best partitioning of the data is chosen using a given *decision rule*. For example, the selected value $\hat{K}$ may be the K with the maximum CVI value.

The performance of various CVIs has previously been compared in the literature. The first comprehensive comparison was conducted by [1], where the Calinski-Harabasz index [2] had the best performance in a Monte Carlo evaluation of 30 indices. This study also introduced the idea that various decision rules could be applied to the same index. The datasets used by the authors had low dimensionality and consisted of distinct, non-overlapping clusters. Chiang and Mirkin [3] compared 9 different indices in a variety of conditions. Their experimental design included, among other factors, varying cluster shapes and cluster overlap. Hartigan's rule [4] achieved the best score in recovering the number of clusters. In a recent study by [5], 30 indices were compared using 6480 different configurations. Cluster separation was introduced in the analysis by varying cluster density and the distance between cluster centroids. The Silhouette width [6] performed the best overall and in the case of strong cluster overlap.

[*] Both authors have contributed equally.

Although several comparison studies have been conducted in the past, most have focused on the case of distinctly separate clusters. However, in real datasets, groups are not always well separated. Studies that have included cluster separation as a factor controlled it using two parameters: distances between centroids and cluster density. Unfortunately, it is difficult to choose an appropriate combination of these two parameters, as different pairs of cluster densities and centroid distances can result in the same degree of cluster overlap. Most previous comparison studies have also ignored the potential of different decision rules to improve the selection of the parameter K.

The objective of this study is to evaluate the impact of cluster separation and high dimensionality on the estimation of partition quality, as well as the impact of the chosen decision rule on the quality of the chosen partition. The J separation index introduced in [7] is used to control cluster separation in the simulated data using a single parameter, enabling more precise control of the degree of cluster separation. In addition, two decision rules are assessed for each index evaluated: the original rule and an alternative. Thus, this study reexamines the question of decision rule choice and demonstrates its importance with respect to partition quality.

Section 2 presents the indices evaluated in this study, Section 3 describes the study protocol and Section 4 presents the main results and concludes the paper.

2 Cluster Validity Indices and Decision Rules

The CVIs studied here (shown in Table 1) were chosen based on their performance in previous comparison studies. All of these CVI functions $I(K)$ involve the calculation of some measure of compactness for each cluster k. We can divide these CVIs into two groups: *global* indices, which only use the points belonging to cluster k to calculate its compactness, and *local* indices, which also consider points belonging to the nearest neighboring cluster.

One example of a local method is the Silhouette width (Sil) [6], which measures partition quality by examining whether points assigned to a specific cluster should not be assigned to its nearest neighbor instead. The recently introduced VCN index (VCN) [8] is a variant of Sil that strives to reduce its computational time. The Davies Bouldin index (DB) and its variation (DB*) [9, 10] also follow a local approach and include measures of cluster compactness and separation.

An emblematic example of a global method is the Calinski-Harabasz index (CH) [2]. It maximizes the ratio of the between-cluster sum of squared distances to the within-cluster sum of squared distances. Another example of a global method is Hartigan's rule (Hart) [4], which evaluates the marginal gain in cluster compactness when the number of clusters is increased. The Gap statistic (Gap) [11] compares cluster compactness for a given partition to the average cluster compactness of the partitions estimated for B random reference distributions.

The original decision rules proposed for choosing $\hat{K}$ (see Table 1) may not always favor the partition that is most similar to the true partition. This idea is tested in the following sections (Sections 3 and 4).

Index	**Type**	$I(K)$ **Formula**	**Original decision rule**
CH [2]	global	$\frac{n-K}{K-1}\frac{\text{Inter}(K)}{\text{Intra}(K)}$	$\hat{K} = \arg\max_K I(K)$
DB [9]	local	$\frac{1}{K}\sum_{k=1}^{K} \max_{C_\ell \in C \setminus C_k} \frac{S(C_\ell)+S(C_k)}{\text{d}(\bar{\mathbf{c}}_\ell, \bar{\mathbf{c}}_k)}$	$\hat{K} = \arg\min_K I(K)$
DB* [10]	local	$\frac{1}{K}\sum_{k=1}^{K} \frac{\max_{C_\ell \in C \setminus C_k} S(C_\ell)+S(C_k)}{\min_{C_\ell \in C \setminus C_k} \text{d}(\bar{\mathbf{c}}_\ell, \bar{\mathbf{c}}_k)}$	$\hat{K} = \arg\min_K I(K)$
Gap [11]	global	$(1/B)\sum_{b=1}^{B} \log(W_K^*(b)) - \log(W_K)$	$\hat{K} = \min K$ s.t. $I(K+1) - I(K) \leq s_{K+1}$
Hart [4]	global	$\left(\frac{\text{Intra}(K)}{\text{Intra}(K+1)} - 1\right)(n-K-1)$	$\hat{K} = \min K$ s.t. $I(K) \leq 10$
Sil [6]	local	$\sum_{k=1}^{K}\sum_{i \in C_k} \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$	$\hat{K} = \arg\max_K I(K)$
VCN [8]	local	$\sum_{k=1}^{K} \frac{bd(C_k)-S(C_k)}{\max\{bd(C_k),S(C_k)\}}$	$\hat{K} = \arg\max_K I(K)$

Table 1: **Cluster Validity Indices**. C is a partition of K clusters, C_k is the k-th cluster in C, $\bar{\mathbf{c}}_k$ is the centroid of cluster k, $\mathbf{x}_i$ is observation i, $\bar{\mathbf{x}}$ is the average of all n observations and d() is the Euclidean distance. Inter(K) is the inter-cluster variation, given by $\sum_{k=1}^{K} |C_k| \text{d}^2(\bar{\mathbf{c}}_k, \bar{\mathbf{x}})$, Intra$(K)$ is the intra-cluster variation, given by $\sum_{k=1}^{K}\sum_{i \in C_k} \text{d}^2(\mathbf{x}_i, \bar{\mathbf{c}}_k)$, $S(C_k)$ is the spread of cluster k, given by $\frac{1}{|C_k|}\sum_{i \in C_k} \text{d}(\mathbf{x}_i, \bar{\mathbf{c}}_k)$, D_k is the sum of squared pairwise distances in cluster k, given by $\sum_{j \neq i | i,j \in C_k} \text{d}^2(\mathbf{x}_i, \mathbf{x}_j)$, $W_K = \sum_{k=1}^{K} \frac{1}{2|C_k|} D_k$, $a(i) = \frac{1}{|C_k|}\sum_{j \neq i | i,j \in C_k} \text{d}(\mathbf{x}_i, \mathbf{x}_j)$, $b(i) = \min_{C_\ell \in C \setminus C_k} \frac{1}{|C_\ell|}\sum_{j \in C_\ell} \text{d}(\mathbf{x}_i, \mathbf{x}_j)$ and $bd(C_k) = \min_{C_\ell \in C \setminus C_k} \frac{1}{|C_k|}\sum_{i \in C_k} \text{d}(\mathbf{x}_i, \bar{\mathbf{c}}_\ell)$. For the Gap statistic, B is the number of random datasets, $W_K^*(b)$ is the value of W_K for the b-th random dataset and s_{K+1} is the standard deviation of $\left[\log(W_{K+1}^*(1)), ..., \log(W_{K+1}^*(B))\right]$.

3 Experimental Protocol

The goals of the following numerical experiments are twofold: (i) to compare the performance of existing CVIs under various simulation conditions, including varying data dimensionality and degrees of cluster separation, and (ii) to assess the impact of decision rule choice on the quality of the chosen partition.

The performance of CVIs is analyzed using simulated data with Gaussian clusters. Each true cluster contains 30 observations, and n is fixed to $K \times 30$. Five simulation parameters are varied (see Table 2). The J separation index introduced in [7] is used to control the degree of separation between clusters. Cluster shape is controlled by fixing the ratio (r_λ) of the smallest to the largest eigenvalue of the variance-covariance matrix for each cluster, where a ratio of 1 corresponds to a spherical cluster and a ratio of 10 corresponds to an elongated cluster. Diagonal variance-covariance matrices are used to generate spherical clusters, whereas elongated clusters are generated using block-diagonal matrices.

This experimental design leads to $3 \times 2 \times 2 \times 2 = 24$ configurations. Ten random datasets are sampled for each configuration and two clustering algorithms (Ward's hierarchical clustering and K-means initialized with K-means++) are applied to each dataset, yielding a total of 480 total partitions per number K of clusters tested. Following a common practice in the literature, the values of K evaluated here are all integers in the range $[2, \sqrt{n}]$.

The seven CVIs from Section 2 are applied to these partitions, and the estimated number of clusters $\hat{K}$ is found using the original decision rule ("Orig") and an alternative decision rule. For the "MaxDec" decision rule,

$$\hat{K} = \arg\max_{K} I(K-1) - I(K), \tag{1}$$

and for the "MaxInc" decision rule,

$$\hat{K} = \arg\max_{K} I(K) - I(K-1). \tag{2}$$

Following the methodology in [12], each $\hat{K}$ is compared to the K resulting in the partition most similar to the true partition. The Adjusted Rand Index (ARI) [13] was chosen to evaluate the similarity between true and estimated partitions. According to [14], the ARI is a good choice for true partitions with large, equal-sized clusters. The criterion used to evaluate the quality of $\hat{K}$ is

$$|\hat{K} - K_{ARI}|, \tag{3}$$

where K_{ARI} is the K with the highest ARI. This criterion should be minimized. For each CVI and decision rule pair, partition choice is also assessed for the case where the true partition is included as a candidate partition, as was done in [15].

Parameter	**Value**
Cluster separation (J)	close ($J = 0.01$), separated ($J = 0.210$), well separated ($J = 0.342$)
Cluster shape	spherical ($r_\lambda = 1$), elongated ($r_\lambda = 10$)
Number of clusters (K_{true})	4, 8
Dimensionality (# of dimensions / n)	low dim (0.5), high dim (1.5)
Clustering algorithm	K-means, Ward's hierarchical clustering

Table 2: Experimental design

4 Results and Conclusions

Table 3 shows the average results for the criterion in Eq. (3). In each column, all methods are compared to the best method using a paired t-test, and the best method is highlighted in bold text in a shaded cell, as well as all methods that are not significantly different from it.

Overall, the best methods for the high-dimensional case are DB* MaxDec, DB MaxDec, Hart MaxDec and Sil MaxInc. In this setting, the only methods with competitive performance for all cluster separation levels are DB* MaxDec for K-means partitions and Hart MaxDec for Ward partitions. For low-dimensional data with low cluster separation, the best method is DB orig (except for Ward partitions, where DB MaxDec and DB* MaxDec perform best). When the true partition is a candidate, Hart MaxDec chooses it the most often (for 63.3% of all datasets), followed by Gap MaxInc (48.3%) and Sil MaxInc (41.9%).

Method	1. Low dim, spherical			2. Low dim, elongated			3. High dim, spherical			4. High dim, elongated		
CH MaxInc	5.1	3.4	5.4	5.8	7.8	3.0	8.7	**1.1**	5.0	5.6	10.2	1.2
CH Orig	13.2	3.6	9.2	13.6	2.2	1.4	3.0	1.6	2.5	1.6	2.0	1.6
DB MaxDec	12.5	3.6	8.2	12.3	**0.8**	**0.8**	**1.6**	**0.7**	1.7	**0.6**	**0.6**	**0.6**
DB Orig	**0.2**	7.8	3.6	**0.4**	11.9	9.9	10.9	6.3	8.2	8.7	11.9	6.4
DB* MaxDec	11.6	3.5	6.5	9.9	**1.4**	1.8	**1.6**	**0.8**	1.7	**0.6**	**0.6**	**0.8**
DB* Orig	2.0	6.4	**2.8**	1.8	9.5	8.3	10.9	6.0	7.8	8.1	11.9	5.5
Gap MaxInc	9.0	**2.2**	5.6	9.3	3.5	1.4	4.2	1.4	**1.4**	1.2	4.8	1.3
Gap Orig	12.6	**1.8**	8.0	11.9	1.3	2.1	2.7	1.6	1.7	**0.8**	1.7	1.6
Hart MaxDec	10.1	2.3	7.0	11.3	2.3	2.0	**2.0**	2.0	1.9	1.8	1.8	2.0
Hart Orig	14.0	4.8	10.0	14.0	2.0	1.6	3.0	1.5	3.0	2.0	2.0	1.5
Sil MaxInc	7.9	**2.0**	4.9	8.8	4.0	2.0	3.9	**1.1**	**1.2**	1.4	5.0	1.3
Sil Orig	12.9	5.3	8.9	12.8	2.0	4.8	6.6	2.1	3.0	2.3	6.8	2.0
VCN MaxInc	5.5	2.7	5.1	7.8	6.1	2.6	5.0	2.3	2.7	2.5	6.4	2.1
VCN Orig	13.7	3.9	9.0	12.8	1.8	2.2	2.7	2.9	3.3	3.1	2.2	3.5

Method	5. Low dim, Kmeans			6. Low dim, Ward			7. High dim, Kmeans			8. High dim, Ward			9. Overall		
CH MaxInc	2.2	5.7	4.6	3.9	9.7	10.2	7.1	5.7	3.0	5.9	5.5	5.8	6.7	5.5	3.7
CH Orig	13.9	4.6	9.4	1.8	12.4	11.2	2.8	1.8	2.4	5.7	3.3	6.7	2.8	1.8	2.3
DB MaxDec	12.6	3.5	8.6	**0.8**	12.7	10.9	**1.8**	**0.8**	1.6	4.2	**2.1**	6.2	**1.5**	**0.8**	**1.6**
DB Orig	**0.4**	7.9	3.7	11.9	**0.2**	**1.3**	9.5	9.0	6.8	7.1	7.9	5.4	10.1	10.1	8.5
DB* MaxDec	9.4	3.6	6.1	**0.7**	12.3	10.4	**1.8**	**1.0**	**1.8**	4.2	**2.1**	6.2	1.9	1.2	2.0
DB* Orig	2.2	6.0	**3.0**	10.4	2.1	2.4	8.9	8.7	6.0	7.1	7.9	5.3	9.1	9.1	7.3
Gap MaxInc	9.2	**2.4**	6.3	3.1	9.4	9.1	**2.5**	2.9	**1.2**	**3.2**	2.9	4.9	2.7	2.8	**1.6**
Gap Orig	12.0	**2.6**	6.9	1.3	11.3	10.2	2.2	1.6	1.7	5.1	2.8	5.5	1.9	1.7	**1.6**
Hart MaxDec	10.7	**3.1**	6.8	1.6	10.9	9.2	**2.1**	2.3	2.0	**3.4**	**2.0**	**4.3**	2.1	2.0	2.0
Hart Orig	14.0	4.8	9.8	2.0	14.0	12.8	3.0	1.8	2.7	6.0	3.8	7.8	3.0	1.8	2.8
Sil MaxInc	7.0	**2.9**	4.9	2.4	10.2	9.3	3.0	3.1	**1.3**	**3.0**	**2.8**	5.1	2.9	3.0	2.0
Sil Orig	13.1	4.9	8.6	2.6	12.2	9.2	4.7	5.0	2.0	4.5	4.0	**4.5**	3.8	3.7	3.2
VCN MaxInc	5.3	3.8	5.0	4.2	9.2	8.4	4.7	4.2	2.4	**3.4**	3.6	**4.4**	4.2	4.3	2.8
VCN Orig	13.6	4.4	8.9	1.9	13.2	11.5	3.5	2.9	3.6	4.8	3.1	**5.2**	2.8	2.2	2.9

Table 3: Average results for $|\hat{K} - K_{ARI}|$. For each numbered column (1-9), the three inner columns represent close, separated and well separated clusters. Within each inner column, results that are not significantly different from the best results are in bold and highlighted in gray. Results in numbered columns 1-4 are an aggregation over all algorithms and values of K_{true}, and results for 5-8 are an aggregation over all cluster shapes and values of K_{true}.

For all indices studied, using an alternative decision rule ("MaxInc" or "MaxDec") generally improves performance compared to the original rule, suggesting that these alternatives are more robust to a variety of data characteristics. These results demonstrate that the choice of decision rule can greatly impact the quality of the chosen partition, which opens the door to further research on the development of new decision rules, as well as CVI and decision rule pairings. Future works could also explore how other factors, such as the presence of sub-clusters or non-uniform cluster sizes, impact partition choice.

Acknowledgements

The second author gratefully acknowledges financial support from the Belgian Fund for Scientific Research (F.R.S.-FNRS, FRIA grant).

References

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

[2] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics: Theory and Methods*, 3(1):1–27, 1974.

[3] M. Chiang and B. Mirkin. Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. *Journal of Classification*, 27(1):3–40, 2010.

[4] J. Hartigan. *Clustering Algorithms.* Wiley & Sons, New York, 1975.

[5] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. Pérez, and I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013.

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

[7] W. Qiu and H. Joe. Generation of random clusters with specified degree of separation. *Journal of Classification*, 23(2):315–334, 2006.

[8] S. Zhou and Z. Xu. A novel internal validity index based on the cluster centre and the nearest neighbour cluster. *Applied Soft Computing*, 71:78–88, 2018.

[9] D. Davies and D. Bouldin. A cluster separation measure. *IEEE transactions on Pattern Analysis and Machine Intelligence*, (2):224–227, 1979.

[10] M. Kim and R.S. Ramakrishna. New indices for cluster validity assessment. *Pattern Recognition Letters*, 26(15):2353–2363, 2005.

[11] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[12] I. Gurrutxaga, J. Muguerza, O. Arbelaitz, J. Pérez, and J. Martín. Towards a standard methodology to evaluate internal cluster validity indices. *Pattern Recognition Letters*, 32(3):505–515, 2011.

[13] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[14] S. Romano, N. Vinh, J. Bailey, and K. Verspoor. Adjusting for chance clustering comparison measures. *The Journal of Machine Learning Research*, 17(1):4635–4666, 2016.

[15] K. Tasdemir and E. Merènyi. A validity index for prototype-based clustering of data sets with complex cluster structures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(4):1039–1053, 2011.